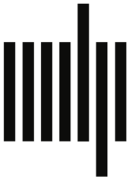
An Introduction to the Event-Related Potential Technique

Steven J. Luck



The MIT Press

From The MIT Press

MITCogNet

This book was set in Melior and Helvetica on 3B2 by Asco Typesetters, Hong Kong.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Luck, Stephen J.
An introduction to the event-related potential technique / Stephen J. Luck.
p. cm. — (Cognitive neuroscience)
Includes bibliographical references and index.
ISBN 0-262-12277-4 (alk. paper) — ISBN 0-262-62196-7 (pbk. : alk. paper)
1. Evoked potentials (Electrophysiology) I. Title. II. Series.
QP376.5.L83 2005
616.8′047547—dc22 2005042810

10 9 8 7 6 5 4 3 2 1

6 Plotting, Measurement, and Analysis

The previous chapters have discussed designing experiments, collecting the data, and applying signal processing procedures; this chapter discusses the final steps in an ERP experiment: plotting the data, measuring the components, and subjecting these measurements to statistical analyses.

Plotting ERP Data

Although it might seem trivial to plot ERP waveforms, I regularly read ERP papers in which the waveforms are plotted in a way that makes it difficult to perceive the key aspects of the waveforms. Thus, I will begin this chapter with a few recommendations about plotting ERP data.

The most important thing to keep in mind is that virtually all ERP papers should include plots of the key ERP waveforms. Given all of the difficulties involved in isolating specific ERP components, it is absolutely necessary to see the waveforms before accepting the validity of a paper's conclusions. Remember that Kramer (1985) showed that ERP experts do a good job of determining the underlying latent components when they see the observed ERP waveforms, and presenting the waveforms is essential for this. In fact, the official publication guidelines of the Society for Psychophysiological Research (SPR) state that ''the presentation of averaged ERP waveforms that illustrate the principal phenomena being reported is mandatory'' (Picton et al., 2000, p. 139).

The SPR guidelines describe several additional elements of ERP figures that you should keep in mind. First, it is almost always a good idea to show the data from multiple electrode sites, spanning

the region of the scalp where the effects of interest were present. This information plays a key role in allowing experts to determine the underlying component structure of the waveform. However, including multiple electrode sites is not as important when you can isolate components by other means (e.g., by performing the subtraction procedure that isolates the lateralized readiness potential).

Second, plots of ERP waveforms should indicate the voltage scale and the time scale in a way that makes it easy for readers to assess amplitudes and latencies. The voltage scale should indicate whether positive or negative is up (and if negative is up, I would recommend stating this in the figure caption). In addition, the electrode sites should be labeled in the figure (if a single site is shown, I often indicate the site in the caption).

Figure 6.1 shows an example of a bad way (panel A) and a good way (panel B) to plot ERP waveforms. The most egregious error in

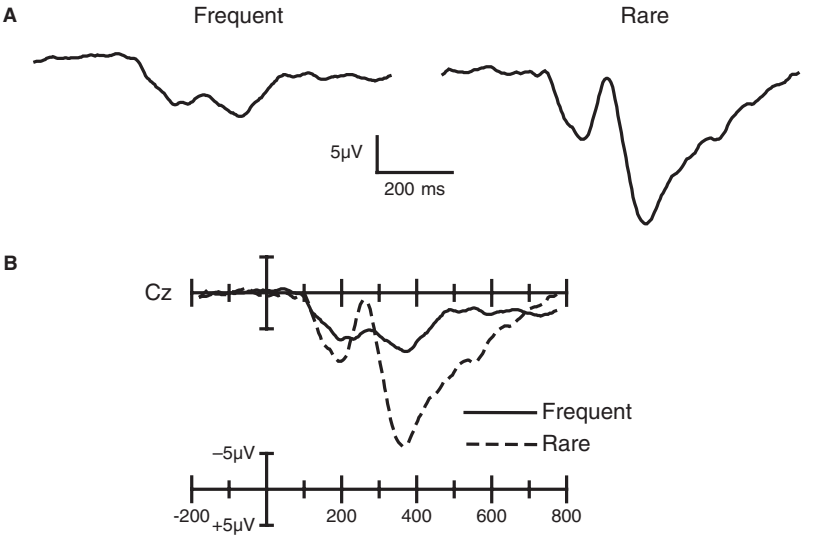


Figure 6.1 Examples of a poor way to plot ERP waveforms (A) and a good way to plot them (B). Negative is plotted upward.

panel A is that there is no X axis line running through zero microvolts to provide an anchor point for the waveforms. It is absolutely essential to include an X axis line that shows the zero microvolt point and the time scale for each waveform. Without it, the reader will find it difficult to determine the amplitude or the latency of the various points in the waveform. A separate scale marker, such as that shown in figure 6.1A, is not enough, because it does not provide a visual reference point. For example, can you tell how much bigger the first component is in the left waveform than in the right waveform? And can you see that this component peaks earlier in the right waveform than in the left waveform?

The second most egregious error in figure 6.1A is that the waveforms are not overlapped. This makes it very difficult for the reader to determine the exact pattern of effects. For example, can you tell exactly when the difference between the waveforms switches polarities? All key effects should be shown by overlapping the relevant waveforms. Figure 6.1A also fails to indicate time zero, and it is impossible to tell which polarity is plotted upward.

Panel B of figure 6.1 solves these problems. It has an X axis running through the waveforms, which are overlapped. It provides additional calibration axis, indicating the latencies (with minor tick marks as well, showing the intermediate latencies). The voltage calibration indicates the polarity of the waveform. The two waveforms are drawn in easy-to-distinguish line types, and the figure provides a legend to indicate what condition each line type represents (this is much better than providing this information in words in the figure caption or main text). The figure also indicates the electrode site at which the waveforms were measured.

It is also essential for the plot to show a significant prestimulus portion of the waveforms. The prestimulus period allows the reader to assess the overall noise level of the recordings and the presence of overlap from the preceding trials. In addition, the apparent amplitude of the ERP peaks will depend on the baseline level, and a stable baseline is therefore essential for accurately characterizing the differences between two or more waveforms.

I almost always provide a 200-ms prestimulus baseline, and I would recommend this for the vast majority of cognitive ERP experiments.

You should be careful when selecting the line types for the ERP waveforms. First, you need to use reasonably thick lines. Figures are usually reduced by a factor of two or three times for publication, making lines that were originally moderately thin become so thin that they are nearly invisible. If someone makes a photocopy of the journal article, portions of the lines may disappear. Of course, you don't want to make the lines so thick that they obscure the experimental effects. I find that a two-point width works well in the majority of cases. In addition, when multiple waveforms overlap, you should make sure that the different lines are easy to differentiate. The best way to do this is with solid, dashed, and dotted lines (or with color, if available, but then photocopies will lose the information). You should avoid using different line thicknesses to differentiate the waveforms; this doesn't work as well, and the thin lines may disappear once the figure has been reduced for publication. In general, you should avoid overlapping more than three waveforms (four if absolutely necessary). With more overlapping waveforms, the figure starts to look like spaghetti.

How many different electrode sites should you show? This depends, in part, on the expected audience. If the paper will be published in an ERP-oriented journal, such as *Psychophysiology*, the readers will expect to see waveforms from a broad selection of electrode sites. If the paper will be published in a journal where ERP studies are not terribly common, such as Journal *of Neuroscience* or *JEP: General*, you should show only the key sites. If you have recorded from a large number of electrodes, it is pointless to include a figure showing the data from all the sites. If a figure contains more than ten to fifteen sites, the individual waveforms will be so small that they will provide very little information. It is much better to show a representative sample of the electrode sites and then use topographic maps to show the distribution of voltage over the scalp.

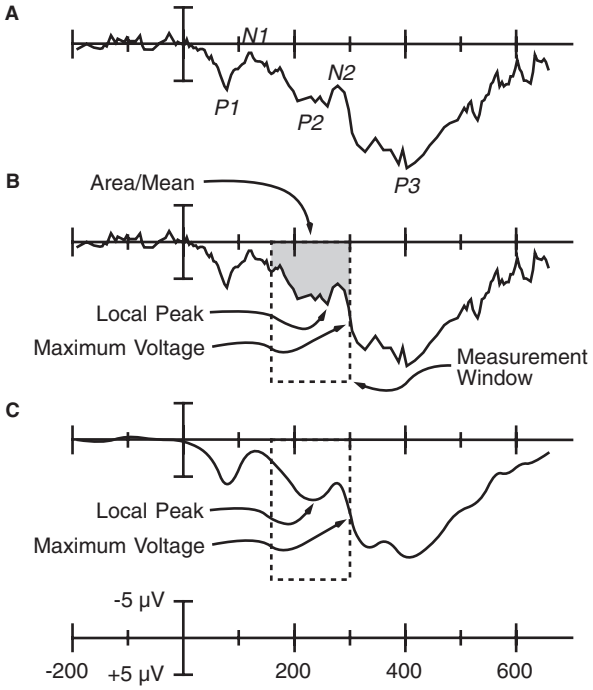


Figure 6.2 (A) ERP waveform with several peaks. (B) ERP waveform, with a measurement window from 150–300 ms poststimulus and several measures. (C) Filtered ERP waveform. Negative is plotted upward.

Measuring ERP Amplitudes

Most ERP studies concentrate on the amplitudes of one or more ERP components, and it is important to measure amplitude accurately. As figure 6.2 illustrates, there are two common ways to measure ERP amplitudes. The most common method is to define a time window and, for each waveform being measured, find the maximum amplitude in that time window. This is called a *peak amplitude* measure. The second most common method is to define a time window and, for each waveform being measured, calculate the mean voltage in that time window. This is called a *mean amplitude* measure. It is also possible to calculate the sum of the voltages at each time point within the measurement window (an *area*

amplitude measure), but this is really just the mean amplitude multiplied by the number of points in the measurement window. For all of these measures, the voltages are typically measured relative to the average voltage in the prestimulus period. Other techniques are sometimes used, but they are beyond the scope of this book.

The goal of these techniques is to provide an accurate measurement of the size of the underlying ERP component with minimal distortion from noise and from other, overlapping components. As I discussed at the beginning of chapter 2, it is very difficult to measure the amplitude and latency directly from a raw ERP waveform without distortion from overlapping components. That chapter also described several strategies for avoiding this problem (e.g., using difference waves), but those strategies often rely on the use of an appropriate measurement approach. And don't forget rule 1 from chapter 2: *Peaks and components are not the same thing. There is nothing special about the point at which the voltage reaches a local maximum.*

Peak Amplitude

Given that the point at which the voltage reaches a local maximum is not special, why should you use peak amplitude to measure the amplitude of an ERP component? In fact, there is no compelling reason to use peak amplitude measures in most ERP experiments (although there might be good reasons to use this measure in some special cases). Moreover, there are several reasons why you shouldn't use peak amplitude in many cases, at least not without some modifications to the basic procedure.

Consider, for example, the ERP waveform shown in panel A of figure 6.2. To measure the peak of the P2 wave, you might use a time window of 150–300 ms. As panel B of figure 6.2 shows, the maximum voltage in this time window occurs at the edge of the time window (300 ms) due to the onset of the P3 wave. Consequently, the amplitude of the P2 wave would be measured at 300

ms, which is not very near the actual P2 peak. Clearly, this is not a good way to measure the P2 wave. You could avoid this problem by using a narrower measurement window, but a fairly wide window is usually necessary because of variations in peak latency across electrode sites, experimental conditions, and subjects. A much better way to avoid this problem is to search the measurement window for the largest point that is surrounded on both sides by smaller points. I call this measure *local peak amplitude*, and I refer to the original method as *simple peak amplitude*. To minimize spurious local peaks caused by high-frequency noise, the local peak should be defined as having a greater voltage than the average of the three to five points on either side rather than the one point on either side.[1]

Local peak amplitude is obviously a better measure than the simple peak amplitude. Strangely, I have never seen anyone explicitly describe this way of measuring peaks. It may be that people use it and just call it peak amplitude (I have done this myself), but it would probably be best if everyone used the term *local peak amplitude* to make it clear exactly how the peaks were measured.

Although local peak amplitude is a better measure than the simple peak amplitude, high-frequency noise can significantly distort both of these measures. In figure 6.2B, for example, the local peak is not in the center of the P2 wave, but is shifted to the right because of a noise deflection. It makes sense that peak amplitude would tend to be noise-prone, because the voltage at a single time point is used to measure a component that lasts for hundreds of milliseconds. Moreover, the peak amplitude measurements will tend to be larger for noisier data and for wider measurement windows, because both of these factors increase the likelihood that a really large noise deflection will occur by chance. Consequently, it is never valid to compare peak amplitudes from averages of different numbers of trials or from time windows of different lengths. I have seen this rule violated in several ERP papers, and it definitely biases the results.

If, for some reason, you still want to measure peak amplitude, you can reduce the effects of high-frequency noise by filtering out the high frequencies before measuring the peak. Panel C of figure 6.2, which shows the waveform in panels B and C after low-pass filtering, illustrates this. In the filtered waveform, the simple peak amplitude again provides a distorted measure, but the local peak provides a reasonably good measure of the amplitude. As discussed in chapter 5, when high frequencies are filtered out, the voltage at each point in the filtered waveform reflects a weighted contribution of the voltages from the nearby time points. Thus, the peak amplitude in a filtered waveform avoids the problem of using the voltage at a single time point to represent a component that lasts hundreds of milliseconds.

Another shortcoming of peak amplitude measures is that they are essentially nonlinear. For example, if you measure the peak amplitude of the P2 wave for several subjects and compute the average of the peak amplitudes, the result will almost always be different from what would you would get by creating a grand average of the single-subject waveforms and measuring the peak amplitude of the P2 wave in this grand average. This may cause a discrepancy between the grand-average waveforms that you present in your figures and the averaged peak amplitude values that you analyze statistically. Similarly, when there is variability in the latency of a component from trial to trial in the raw EEG data, the peak amplitude in the averaged ERP waveform will be smaller than the single-trial amplitudes (see panels G and H of figure 2.1 in chapter 2). If latency variability is greater in one condition than in another, the peak amplitudes will differ between conditions even if there is no difference between conditions in the single-trial peak amplitudes.

To summarize, peak amplitude measures have four serious shortcomings: (1) when the maximum voltage in the measurement window is measured, the rising or falling edge of an overlapping component at the border of the measurement window may be measured rather than the desired peak; (2) peak amplitude uses a

single point to represent a component that may last hundreds of milliseconds, making it sensitive to noise; (3) peak amplitude will be artificially increased if the noise level is higher (due, for example, to a smaller number of trials) or if the measurement interval is longer; (4) peak amplitude is a nonlinear measure that may not correspond well with grand-average ERP waveforms or with single-trial peaks.

It might be tempting to argue that peak amplitude is a useful measure because it takes into account variations in latency among conditions, electrode sites, and subjects. After all, you wouldn't want to measure the amplitude of a component at the same time point at all electrode sites if the component peaks later at some sites than at others, would you? This argument is fallacious and points to a more abstract reason for avoiding peak amplitude measures. Specifically, peak amplitude measures implicitly encourage the mistaken view that peaks and components are the same thing and that there is something special about the point at which the voltage reaches its maximum value.

For example, the fact that the voltage reaches its maximum value at different times for different electrode sites is unlikely to reflect differences in the latency of the underlying component at different electrode sites. A component, as typically defined, is a single brain process that influences the recorded voltage simultaneously at all electrode sites, and it cannot have a different latency at different sites. As described in chapter 2 (see especially figure 2.1), the timing of a peak in the observed ERP waveform will vary as a function of the relative amplitudes of the various overlapping components, and this is the reason why peak latencies vary across electrode sites. Indeed, it would be rather strange to measure the same component at different times for the different electrode sites. Similarly, peak latency differences between subjects are just as likely to be due to differences in the relative amplitudes of the various underlying components as differences in the latency of the component of interest. In contrast, the latency of the underlying component

may indeed vary considerably across different experimental conditions. But in this case, it will be nearly impossible to measure the component without significant distortion from other overlapping components, because amplitude and latency measures are often confounded (see chapter 2). Thus, although peak amplitude measures might seem less influenced by changes in component timing, this is really an illusory advantage.

Mean Amplitude

Mean amplitude has several advantages over peak amplitude. First, you can use a narrower measurement window because it doesn't matter if the maximum amplitude falls outside of this window for some electrode sites or some subjects. In fact, the narrower the window, the less influence overlapping components will have on the measurements. As an example, a measurement window of 200–250 ms would be appropriate for measuring the mean amplitude of the P2 component in the waveform shown in figure 6.2, compared to a window of 150–300 ms that would be appropriate for a peak amplitude measure.

A second advantage of mean amplitude measures is that they tend to be less sensitive to high-frequency noise than are peak amplitude measures, because a range of time points is used rather than a single time point. For this reason, you don't want to make your measurement windows too narrow (< 40 ms or so), even though a narrow window is useful for mitigating the effects of overlapping components. You should also note that, although filtering can reduce the effects of high-frequency noise in peak amplitude measures, there is no advantage to low-pass filtering when measuring mean amplitude. By definition, mean amplitude includes the voltages from multiple nearby time points, and this is exactly what low-pass filtering does. In fact, filtering the data before measuring the mean amplitude in a given measurement window is the same thing as measuring the mean amplitude from an unfiltered waveform using a wider measurement window.

A third advantage of mean amplitude measures is that they do not become biased when the noise level increases or when one uses a longer measurement window. In other words, the variance may change, but the expected value is independent of these factors. Consequently, it is legitimate to compare mean amplitude measurements from waveforms based on different numbers of trials, whereas this is not legitimate for peak amplitude measurements.

A fourth advantage is that mean amplitude is a linear measure. That is, if you measure the mean amplitude of a component from each subject, the mean of these measures will be equal to measuring the mean amplitude of the component from the grand-average waveform. This makes it possible for you to compare directly your grand-average waveforms with the means from your statistical analyses. This same principle also applies to the process of averaging together the single-trial EEG data to form averaged ERP waveforms. That is, the mean amplitude measured from a subject's averaged ERP waveform will be the same as the average of the mean amplitudes measured from the single-trial EEG data.

Although mean amplitude has several advantages over peak amplitude, it is not a panacea. In particular, mean amplitude is still quite sensitive to the problem of overlapping components and can lead to spurious results if the latency of a component varies across conditions. In addition, there is not always an a priori reason to select a particular measurement window, and this can encourage fishing for significant results by trying different windows. There are some more sophisticated ways of measuring the amplitude of a component (e.g., dipole source modeling, ICA, PCA, etc.), but these methods are based on a variety of difficult-to-assess assumptions and are beyond the scope of this book (for more information on alternative measures, see Coles et al., 1986). Thus, I would generally recommend using mean amplitude measures in conjunction with the rules and strategies for avoiding the problem of overlapping components discussed in chapter 2.

Baselines

Whether you are measuring mean amplitude or peak amplitude, you will be implicitly or explicitly subtracting a voltage—usually the average prestimulus voltage—that represents the baseline or zero point. Any noise in the baseline will therefore add noise to your measures, so it is important to select an appropriate baseline. When you are measuring stimulus-locked averages, I would recommend that you use the average voltage in the 200 ms before stimulus onset as the baseline. A 100-ms baseline is common, and although it's not as good as 200 ms, it's usually acceptable. If you use less than 100 ms, it is likely that you will be adding noise to your measures.

The prestimulus interval is usually used as the baseline because it is assumed that the voltage in this period is unaffected by the stimulus. Although it is true that the time-locking stimulus cannot influence the prestimulus period, it is important to realize that the processing does not always begin after the onset of a stimulus. If the interval between stimuli is relatively short (< 2 s), the late potentials from the previous stimulus may not have completely diminished by the time of the time-locking stimulus and may therefore contribute to the baseline voltage. In addition, regardless of the length of the interstimulus interval, the prestimulus voltage may be influenced by preparatory processes (in fact, the effects of preparatory processes are more visible when the interstimulus interval is long). For these reasons, the prestimulus baseline rarely provides a perfectly neutral baseline, and in many experiments it is clear that the voltage slopes upward or downward during the prestimulus interval because of these factors. This can be a major problem if the prestimulus activity differs across experimental conditions, because any difference in measured amplitudes between conditions might reflect prestimulus differences rather than poststimulus differences (see chapter 4, and Woldorff, 1988).

Even when the prestimulus activity does not differ across conditions, it is important to recognize that the prestimulus interval is usually not completely neutral. For example, scalp distributions can be significantly distorted if significant activity is present in the

prestimulus interval and then fades by the time of the component being measured. The only real solution is to keep in mind that the measured voltage reflects the difference between the amplitude in the measurement window and the amplitude in the prestimulus period (just as it also reflects the difference between the active and reference electrodes).

Baselines become much more complicated when using response-locked averages, because the activity that precedes the response is often as large as or larger than the activity that follows the response. One solution to this problem is to use as the baseline a period of time that precedes the response by enough so that it always precedes the stimulus. Another solution is to use the average voltage of the entire averaging epoch as the baseline. As far as I am aware, there is no single best solution, other than thinking carefully about whatever baseline period you use to make sure that it isn't distorting your results.

I should also mention that investigators sometimes compute *peak-to-peak* amplitudes that consist of the difference in amplitude between two successive peaks of opposite polarity. For example, the amplitude of the visual N1 wave is sometimes measured as the difference in amplitude between the N1 peak and the preceding P1 peak (this could also be done with mean amplitudes, although this is much less common). This approach is sometimes taken if the components are assumed to overlap in time such that the earlier component distorts the amplitude of the later component. This may be a reasonable approach, as long as you keep in mind exactly what you are measuring.

Measuring ERP Latencies

Peak Latency

ERP latencies are usually measured with *peak latency* measures that find the maximum amplitude within a time window and use

the latency of this peak as a measure of the latency of the underlying component. All of the shortcomings of peak amplitude measures also have analogs in peak latency measures. First, when the measurement window includes the rising or falling edge of a larger component, the maximum voltage will be at the border of the window (see figure 6.2B). Fortunately, this problem can be solved by using a *local peak latency* measure in which a point is not considered a peak unless the three to five points on each side of it have smaller values.

A second shortcoming of peak latency measures is that, like peak amplitude measures, they are highly sensitive to high-frequency noise. If a peak is rather broad and flat, high-frequency noise may cause the maximum voltage to be very far away from the middle of the peak (see the local peak in figure 6.2B). In fact, this is probably an even more significant problem for peak latency than for peak amplitude, because a noise-related peak may be far in time from the true peak but close to the true peak's amplitude. As in the case of peak amplitude, you can mitigate this problem to some extent by filtering out the high frequencies in the waveform (see the local peak in figure 6.2C).

A third shortcoming is that peak latency, like peak amplitude, will change systematically as noise levels increase. Specifically, as the noise increases, the average peak latency will tend to be nearer to the center of the measurement window. To understand why this is true, imagine an ERP waveform that is entirely composed of random noise. If you measure the peak latency between 200 and 400 ms, the peak latency on any given trial is equally likely to be at any value in this range, and the average will therefore be at the center of the range. If there is a signal as well as noise, then the average will tend to be somewhere between the actual peak and the center of the range.

A fourth shortcoming of peak latency is that it is nonlinear. In other words, the peak latency measured from a grand average will not usually be the same as the average of the peak latencies that were measured from the single-subject waveforms. Similarly,

the peak latency measured from a single-subject average will not usually be the same as the average of the single-trial peak latencies.

A fifth shortcoming is that peak latency measures, like peak amplitude measures, implicitly encourage the mistaken view that peaks and components are the same thing and that there is something special about the point at which the voltage reaches its maximum value. The latency at which the voltage reaches its maximum value depends greatly on the nature of the overlapping components and the waveshape of the component of interest, so peak latency bears no special relationship to the timing of an ERP component.

Although peak latency has many shortcomings, there just aren't many good alternatives, and so it is often the best measure. When measuring peak latency, you should take the following precautions: (1) filter out the high-frequency noise in the waveforms; (2) use a local peak measure rather than an absolute peak measure; (3) make sure that the waveforms being compared have similar noise levels; and (4) keep in mind that peak latency is a coarse and nonlinear measure of a component's timing.

Fractional Area Latency

Under some conditions, it is possible to avoid some of the shortcomings of peak latency measures by using *fractional area latency* measures, which are analogous to mean amplitude measures. Fractional area measures work by computing the area under the ERP waveform over a given latency range and then finding the time point that divides that area into a prespecified fraction (this approach was apparently first used by Hansen & Hillyard, 1980). Typically the fraction will be a half, in which case this would be called a *50 percent area latency* measure. Figure 6.3A shows an example of this. The measurement window in this figure is 300–600 ms, and the area under the curve in this time window is divided at 432 ms into two regions of equal area. Thus, the 50 percent area latency is 432 ms.

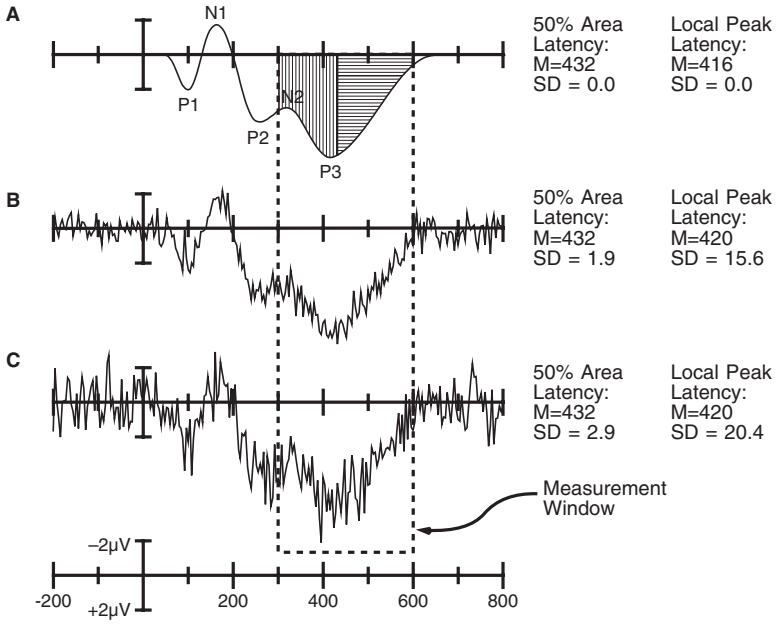


Figure 6.3 Application of 50 percent area latency and local peak latency measures to a noise-free ERP waveform (A), an ERP waveform with a moderate amount of noise (B), and an ERP waveform with significant noise (C). In each case, measurements were obtained from 100 waveforms, and the mean (M) and standard deviation (SD) across these 100 measures are shown. Negative is plotted upward.

The latency value estimated in this manner will depend quite a bit on the measurement window chosen. For example, if the measurement window for the waveform shown in figure 6.3A was shortened to 300–500 ms rather than 300–600 ms, an earlier 50 percent area latency value would have been computed. Consequently, this measure is not appropriate for estimating the absolute latency of a component unless the measurement window includes the entire component and no other overlapping components are present. However, peak latency also provides a poor measure of absolute latency because it is highly distorted by overlapping components and is relatively insensitive to changes in waveshape. Fortunately, in the vast majority of experiments, you don't really care

about the absolute latency of a component and are instead concerned with the relative latencies in two different conditions. Consequently, when you are deciding how to measure the latencies in an experiment, you should base your decision on the ability of a measure to accurately characterize the difference in latency between two conditions. In many cases, the 50 percent area latency measure can accurately quantify latency differences even when the measurement window does not contain the entire component and when there is some overlap from other components.

One advantage of the 50 percent area latency measure is that it is less sensitive to noise. To demonstrate this, I added random (Gaussian) noise the waveform shown in figure 6.3B and then measured the 50 percent area latency and the local peak latency of the P3 wave. I did this a hundred times for each of two noise levels, making it possible to estimate the variability of the measures. When the noise level was 0.5 μV, the standard deviation of the peak latency measure over the hundred measurements was 15.6 ms, whereas the standard deviation of the 50 percent area latency measure was only 1.9 ms. This is shown in figure 6.3B, which shows the waveform from one trial of the simulation (the basic waveform was the same on each of the hundred trials, but the random noise differed from trial to trial). When the noise level was increased to 1.0 μV (figure 6.3C), the standard deviation of the peak latency measure was 20.4 ms, whereas the standard deviation of the 50 percent area latency measure was only 2.9 ms. The variability in peak latency measures can be greatly decreased by filtering the data, and a fair test of peak latency should be done with filtered data. When the waveforms were filtered by convolving them with a Gaussian impulse response function (SD = 30 ms, half-amplitude cutoff at 6 Hz), the standard deviations of the peak latency measures dropped to 3.3 ms and 6.1 ms for noise levels of 0.5 and 1.0 μV, respectively. Thus, even when the data were filtered, the standard deviation of the peak latency measure was approximately twice as large as the standard deviation of the 50 percent area latency measure.

The 50 percent area latency measure has several other advantages as well. First, it is a sensible way to measure the timing of a component that doesn't have a distinct peak or has multiple peaks. Second, it has the same expected value irrespective of the noise level of the data. Third, the 50 percent area latency is linear, so the mean of your measures will correspond well with what you see in the grand-average waveforms. Finally, 50 percent area latency can be related to reaction time more directly than peak latency, as described in the next section.

Although the 50 percent area latency measure has several advantages over peak latency, it also has a significant disadvantage. Specifically, it can produce very distorted results when the latency range does not encompass most of the ERP component of interest or when the latency range includes large contributions from multiple ERP components. Unfortunately, this precludes the use of the 50 percent area latency measure in a large proportion of experiments. It is useful primarily for late, large components such as P3 and N400, and primarily when the component of interest can be isolated by means of a difference wave. In my own research, for example, I have used the 50 percent area latency measure in only two studies (Luck, 1998b; Luck & Hillyard, 1990). However, these were the two studies in which I was most interested in measuring ERP latencies and comparing the results with RT. Thus, 50 percent area latency can be a very useful measure, but it is somewhat limited and works best in experiments that have been optimized to isolate a specific ERP component.

I would like to end this section by emphasizing that it is very difficult to measure ERP latencies. The peak latency measure has several problems, and although the 50 percent area latency measure has some advantages over peak latency, it is appropriate only under a restricted set of conditions. Moreover, the sophisticated PCA and ICA techniques that have been developed for measuring ERP amplitudes cannot be used to measure latencies (and generally assume that latencies are constant). Thus, you must be extremely careful when interpreting latency measures.

Comparing ERP Latencies with Reaction Times

In many studies, it is useful to compare the size of an ERP latency effect to the size of a reaction time (RT) effect. This seems straightforward, but it is actually quite difficult. The problem is that RT is usually summarized as the mean across many trials, whereas a peak in an ERP waveform is more closely related to the peak (i.e., the mode) of the distribution of single-trial latencies, and 50 percent area latency is more closely related to the median of this distribution. In many experiments, RT differences are driven by changes in the tail of the RT distribution, which influences the mean much more than the mode or the median. Consequently, ERP latency effects are commonly smaller than mean RT effects.

To make this clearer, figure 6.4A shows the probability distribution of RT in two conditions of a hypothetical experiment, which we'll call the *easy* and *difficult* conditions. Each point represents the probability of an RT occurring with $\pm$15 ms of that time (that is, the figure is a histogram with a bin width of 30 ms). As is typical, the RT distributions are skewed, with a right-side tail extending out to long RTs, and much of the RT difference between the conditions is due to a change in the probability of relatively long RTs rather than a pure shift in the RT distribution. Imagine that the P3 wave in this experiment is precisely time-locked to the response, always peaking 150 ms after the RT. Consequently, the P3 wave will occur at different times on different trials, with a probability distribution that is shaped just like the RT distribution from the same condition (but shifted rightward by 150 ms). Imagine further that the earlier components are time-locked to the stimulus rather than the response. Figure 6.4B shows the resulting averaged ERP waveforms for these two conditions.

Because most of the RTs occur within a fairly narrow time range in the easy condition, most of the single-trial P3s will also occur within a narrow range, causing the peak of the averaged ERP waveform to occur approximately 150 ms after the peak of the RT distribution (overlap from the other components will influence the precise latency of the peak). Some of the single-trial RTs occur

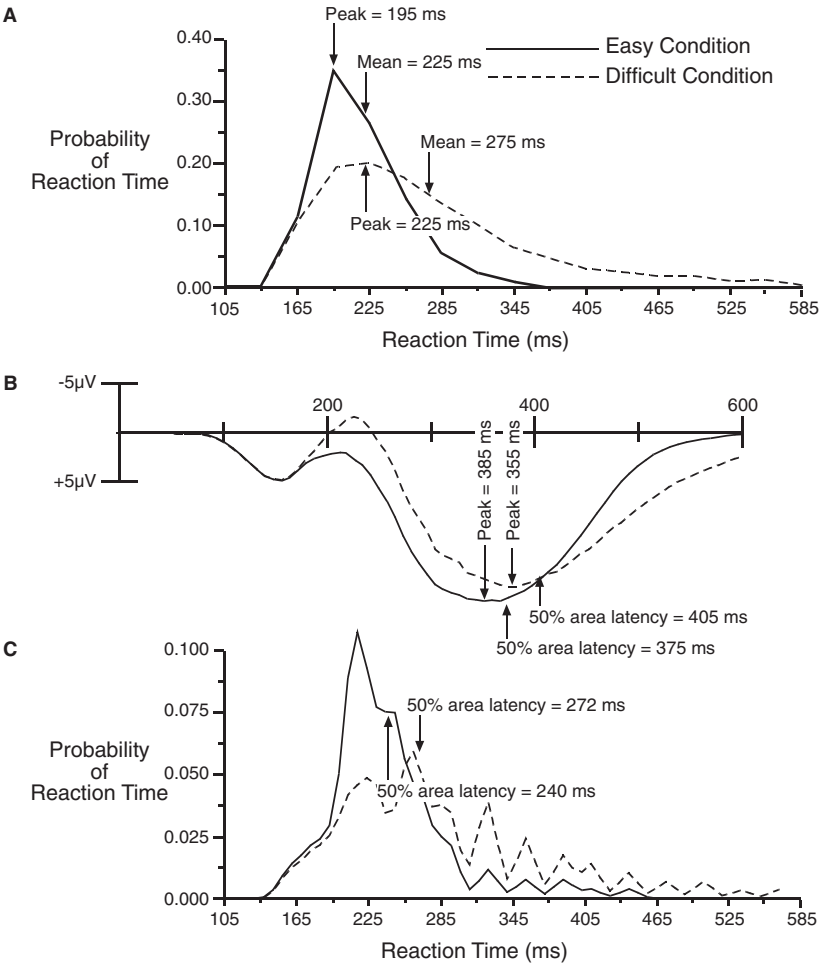


Figure 6.4 (A) Histogram showing the probability of a response occurring in various time bins (bin width = 30 ms) in an easy condition and a difficult condition of a hypothetical experiment. (B) ERP waveforms that would be produced in this hypothetical experiment if the early components were insensitive to reaction time and the P3 wave was perfectly time-locked to the responses. Negative is plotted upward. (C) Response density waveforms showing the probability of a response occurring at any given moment in time.

at longer latencies, but they are sufficiently infrequent that they don't have much influence on the peak P3 latency in the averaged waveform.

The mean RT is 50 ms later in the difficult condition than in the easy condition. However, because much of the RT effect consists of an increase in long RTs, the peak of the RT distribution is only 30 ms later in condition B than in condition A. Because the peak of the P3 wave in the averaged ERP waveform is tied closely to the peak of the RT distribution, peak P3 latency is also 30 ms later in the difficult condition than in the easy condition. Thus, the peak latency of the P3 wave changes in a manner that reflects changes in the peak (mode) of the RT distribution rather than its mean. Thus, when RT effects consist largely of increases in the tail of the distribution rather than a shift of the whole distribution, changes in peak latency will usually be smaller than changes in mean RT, even if the component and the response are influenced by the experimental manipulation in exactly the same way. Consequently, you should never compare the magnitude of an ERP peak latency effect to a mean RT effect unless you know that the RT effect is simply a rightward shift in the RT distribution (which you can verify by plotting the distributions, as in figure 5.4A).

How, then, can you compare RT effects to ERP latency effects? The answer is you must measure them in the same way. One way to achieve this would be to use a peak latency measure for both the ERPs and the RTs (using the probability distribution to find the peak RT). However, the peak of the RT distribution is unlikely to reflect the RT effects very well, because the effects are often driven by changes in the longer RTs. In addition, RT distributions are typically very noisy unless they are computed from thousands of RTs, so it is difficult to get a stable measure of the peak of the RT distribution in most experiments. Moreover, the latency of a peak measured from an ERP waveform is not a particularly good measure of the timing of the underlying component, as discussed earlier in this chapter.

The 50 percent area latency measure provides the time point that bisects the area of the waveform, and this is similar to the median RT, which is the point separating the fastest half of the RTs from the slowest half. Indeed, I once conducted a study which compared the 50 percent area latency of the P3 wave with median RT, and the results were quite good (Luck, 1998b). However, 50 percent area latency is not quite analogous to median RT, because median RT does not take into account the precise values of the RTs above and below the median. For example, a median RT of 300 ms would be obtained for an RT distribution in which half of the values were between 200 and 300 ms and the other half were between 300 and 400 ms, and the same median RT would be obtained if half of the RTs were between 290 and 300 ms and the other half were between 300 and 5,000 ms.

Thus, we need a measure of RT that represents the point that bisects the *area* of the RT distribution, just as the 50 percent area latency measure represents the point that bisects the area under the ERP waveform. The problem is that RTs are discrete, instantaneous events, and this is a problem for area measures. One way to achieve a curve with a measurable area would be to compute histograms of RT probability, as in figure 6.4A. However, these histograms are not really continuous, so they do not provide a perfect measure of area. Researchers who conduct single-unit recordings have to deal with this same problem, because spikes are also treated as discrete, instantaneous events. To create a continuous waveform from a set of spikes, they have developed *spike density waveforms* (see, e.g. Szücs, 1998). In these waveforms, each spike is replaced by a continuous gaussian function, peaking at the time of the spike, and the gaussians are summed together. The same thing can be done for reaction times, creating *response density waveforms*.

Figure 6.4C shows an example of such a waveform. Each individual response that contributed to the histogram in figure 6.4A was replaced by a Gaussian function with a standard deviation of approximately 8 ms, and the average across trials was then com-

puted. The 50 percent area latency was then computed, just as it was for the ERPs. The resulting latencies were 240 ms for the easy condition and 272 ms for the difficult condition, and the 32-ms difference between these latencies compares favorably with the 30-ms effect observed in the 50 percent area latency measures of the ERP waveform.

I don't know of anyone who has tried this approach, and I'm sure it has limitations (I described some of the limitations of the 50 percent area latency measure toward the end of the previous section). However, this approach to comparing RT effects with ERP latency effects is certainly better than comparing peak ERP latencies with mean RTs.

Onset Latency

In many experiments, it is useful to know if the onset time of an ERP component varies across conditions. Unfortunately, onset latency is quite difficult to measure. When the onset of the component of interest is overlapped by other ERP components, there is no way to accurately measure the onset time directly from the waveforms. To measure the onset time, it is necessary to somehow remove the overlapping components. The most common way to do this is to form difference waves that isolate the component of interest (such as the difference waves used to extract the lateralized readiness potential, as described in chapter 2). More sophisticated approaches, such as dipole source analysis and independent components analysis may also be used to isolate an ERP component.

Once a component has been isolated, it is still quite difficult to measure its onset time. The main reason for this is that the onset is a point at which the component's amplitude is, by definition, at or near zero, and the signal-to-noise ratio is also at or near zero. Thus, any noise in the waveform will obscure the actual onset time. Researchers have used several different methods to measure onset time over the years; I'll describe a few of the most common methods.

One simple approach is to plot the waveforms for each subject on paper and have a naïve individual use a straight-edge to extrapolate the waveform to zero μV and then record the latency at that point. This can be time consuming, and it assumes that the component can be accurately characterized as a linearly increasing waveform. Despite the fact that this approach is somewhat subjective, it can work quite well because it takes advantage of the sophisticated heuristics the human visual system uses. But it is also prone to bias, so the person who determines the latencies must be blind to the experimental conditions for each waveform.

Another approach is to use the fractional area latency measure described previously in this chapter. Instead of searching for the point that divides the waveform into two regions of equal area, it is possible to search for the point at which 25 percent of the area has occurred (see, e.g., Hansen & Hillyard, 1980). This can work well, but it can be influenced by portions of the component that occur long after the onset of the component. Imagine that waveforms A and B have the same onset time, but waveform B grows larger than waveform A at a long latency. Waveform B will therefore have a greater overall area than waveform A, and it will take longer to reach 25 percent of the area for waveform B.

Another approach is to find the time at which the waveform's amplitude exceeds the value expected by chance. The variation in prestimulus voltage can be used to assess the amplitude required to exceed chance, and the latency for a given waveform is the time at which this amplitude is first reached (for details, see Miller, Patterson, & Ulrich, 1998; Osman et al., 1992). Unfortunately, this method is highly dependent on the noise level, which may vary considerably across subjects and conditions.

A related approach is to find the time at which two conditions become significantly different from each other across subjects, which provides the onset latency of the experimental effect. The simplest way to achieve this is to perform a t-test of the two conditions at each time point and find the first point at which the corresponding p-value is less than .05. The problem with this is that

many comparisons are being made, making it likely that a spurious value will be found. One could use an adjustment for multiple comparisons, but this usually leads to a highly conservative test that will become significant much later than the actual onset time of the effect. A compromise approach is to find the first time point that meets two criteria: (1) the p-value is less than .05, and (2) the p-values for the subsequent N points are also less than .05 (where N is usually in the range of 3–10). The idea is that it is unlikely to have several spurious values in a row, so this approach won't pick up spurious values. That is fairly reasonable, but it assumes that the noise at one time point is independent of the noise at adjacent time points. This is generally not true, and it is definitely false if the data have not been low-pass filtered. As discussed in chapter 5, low-pass filtering spreads the voltage out in time, and a noise blip at one point in time will therefore be spread to nearby time points. However, if the data have not been extensively low-pass filtered, and N is sufficiently high (e.g., 100 ms), this approach will probably work well.

Miller, Patterson, and Ulrich (1998) have developed an excellent approach for measuring the difference in the onset latency of the lateralized readiness potential between two conditions, and this approach should also be useful for other components that can be isolated from the rest of the ERP waveform. In this approach, one uses the grand-average waveforms from two conditions and finds the point at which the voltage reaches a particular level (e.g., 50 percent of the peak amplitude of the waveform). The difference between the two conditions provides an estimate of the difference in onset times. Because this estimate is based on the point at which a reasonably large amplitude has been reached in the grand average, it is relatively insensitive to noise. One can then test the statistical significance of this estimate by using the *jackknife technique* to assess the standard error of the estimate (see Miller, Patterson, & Ulrich, 1998 for details). This is a well-reasoned technique, and Miller and colleagues (1998) performed a series of simulations to demonstrate that it works well under realistic conditions.

Statistical Analyses

Once you have collected ERP waveforms from a sample of subjects and obtained amplitude and latency measures, it is time to perform some statistical analyses to see whether your effects are significant. In the large majority of cognitive ERP experiments, the investigators are looking for a main effect or an interaction in a completely crossed factorial design, and ANOVA is therefore the dominant statistical approach. Consequently, this is the only approach I will describe. Other approaches are sometimes useful, but they are beyond the scope of this book.

Before I begin describing how to use ANOVAs to analyze ERP data, I would like to make it clear that I consider statistics to be a necessary evil. We often treat the .05 alpha level as being somehow magical, with experimental effects that fall below $p < .05$ as being ''real'' and effects that fall above $p < .05$ as being nonexistent. This is, of course, quite ridiculous. After all, if $p = .06$, there is only a 6 percent probability that the effect was due to chance, which is still rather low. Moreover, the assumptions of ANOVA are violated by almost every ERP experiment, so the p-values that we get are only approximations of the actual probability of a type I error. On the other hand, when examining a set of experimental effects that are somewhat weak, you need a criterion for deciding whether to believe them or not, and a p-value is usually better than nothing.

I have two specific recommendations for avoiding the problems associated with conventional statistics. First, whenever possible, try to design your experiments so that the experimental effects are quite large relative to the noise level and the p-values are very low (.01 or better). That way, you won't have to worry about one ''bad'' subject keeping your p-values from reaching the magical .05 criterion. Moreover, when your effects are large relative to the noise level, you can have some faith in the details of the results that are not being analyzed statistically (e.g., the onset and offset times of the experimental effects). My second suggestion is to use the one statistic that cannot fail, namely replication (see box 6.1). If you

Box 6.1 The Best Statistic

Replication is the best statistic. I learned this when I was a graduate student in the Hillyard lab, although no one ever said it aloud. I frequently say it aloud to my students. Replication does not depend on assumptions about normality, sphericity, or independence. Replication is not distorted by outliers. Replication is a cornerstone of science. Replication is the best statistic.

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

keep getting the same effect in experiment after experiment, then it's a real effect. If you get the effect in only about half of your experiments, then it's a weak effect and you should probably figure out a way to make it stronger so that you can study it more easily.

The Standard Approach

To explain the standard approach to analyzing ERP data with ANOVAs, I will describe the analyses that we conducted for the experiment described briefly near the beginning of chapter 1. In this experiment, we obtained recordings from lateral and midline electrode sites at frontal, central, and parietal locations (i.e., F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4). Subjects saw a sequence of Xs and Os, pressing one button for Xs and another for Os. On some trial blocks, X occurred frequently ($p = .75$) and O occurred infrequently ($p = .25$), and on other blocks this was reversed. We also manipulated the difficulty of the X/O discrimination by

varying the brightness of the stimuli. The experiment was focused on the P3 wave, but I will also discuss analyses of the P2 and N2 components.

Before measuring the amplitudes and latencies of these components, we first combined the data from the Xs and Os so that we had one waveform for the improbable stimuli and one for the probable stimuli. We did this for the simple reason that we didn't care if there were any differences between Xs and Os per se, and collapsing across them reduced the number of factors in the ANOVAs. The more factors are used in an ANOVA, the more individual p-values will be calculated, and the greater is the chance that one of them will be less than .05 due to chance (this is called an increase in *experimentwise* error). By collapsing the data across irrelevant factors, you can avoid this problem (and avoid having to come up with an explanation for a weird five-way interaction that is probably spurious).

Figure 6.5 illustrates the results of this experiment, showing the ERP waveforms recorded at Fz, Cz, and Pz. From this figure, it is clear that the P2, N2, and P3 waves were larger for the rare stimuli than for the frequent stimuli, especially when the stimuli were bright. Thus, for the amplitude of each component, we would expect to see a significant main effect of stimulus probability and a significant probability × brightness interaction.

The analysis of P3 amplitude is relatively straightforward. We measured P3 amplitude as the mean amplitude between 300 and 800 ms at each of the nine electrode sites and entered these data into a within-subjects ANOVA with four factors: stimulus probability (frequent vs. rare), stimulus brightness (bright vs. dim), anterior-posterior electrode position (frontal, central, or parietal), and left-right electrode position (left hemisphere, midline, or right hemisphere). We could have used a single factor for the electrode sites, with nine levels, but it is usually more informative to divide the electrodes into separate factors representing different spatial dimensions. Consistent with the waveforms shown in figure 6.5, this ANOVA yielded a highly significant main effect of stimulus

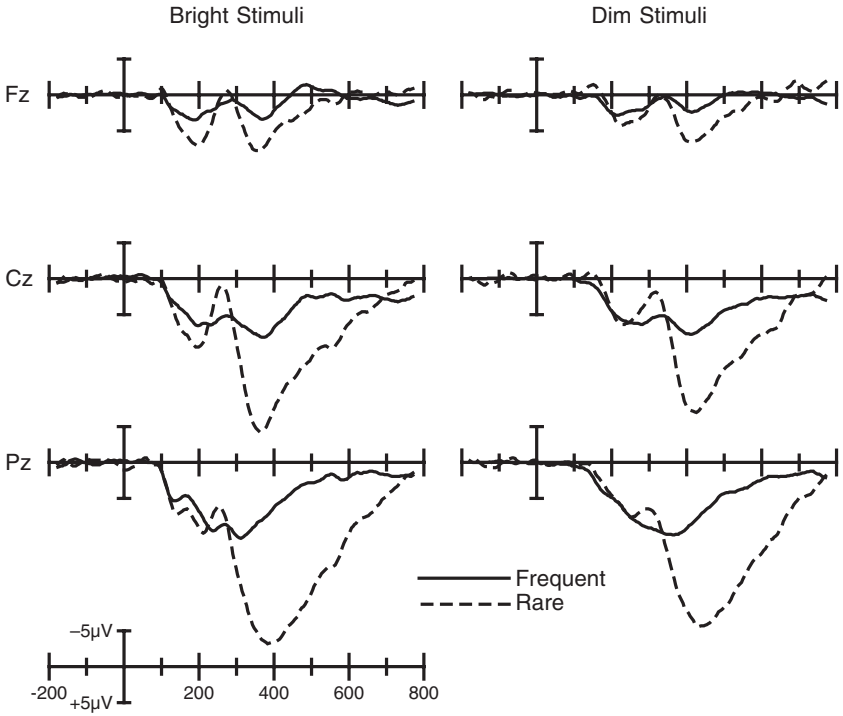


Figure 6.5 Grand-average ERPs for frequent and rare stimuli in the bright and dim conditions. Negative is plotted upward.

probability, $F(1, 9) = 95.48$, $p < .001$, and a significant interaction between probability and brightness, $F(1, 9) = 11.66$, $p < .01$.

Some investigators perform a separate ANOVA for each electrode site (or each left-midline-right set) rather than performing a single ANOVA with electrode site as a factor. Although there may be advantages to this approach, it is likely to increase both the probability of a type I error (incorrectly rejecting the null hypothesis) and the probability of a type II error (incorrectly accepting the null hypothesis). Type I errors will be increased because more p-values must be computed when a separate ANOVA is performed for each electrode, leading to a greater probability that a spurious effect will reach the .05 level. Type II errors will be increased

because a small effect may fail to reach significance at any individual site, whereas the presence of this effect at multiple sites may be enough to make the effect significant when all of the sites contribute to the analysis.

Even when a single ANOVA includes multiple electrode sites, it is usually best not to include measurements from electrode sites spanning the entire scalp. Instead, it is usually best to measure and analyze an ERP component only at sites where the component is actually present. Otherwise, the sites where the component is absent may add noise to the analyses, or the presence of other components at those sites may distort the results. In addition, it is sometimes useful to analyze only the sites at which the component of interest is large *and* other components are relatively small so that they do not distort measurements of the component of interest. In the present study, for example, we used all nine sites for analyzing the P3 wave, which was much larger than the other components, but we restricted the P2 analyses to the frontal sites, where the P2 effects were large but the N2 and P3 waves were relatively small. However, when you are trying to draw conclusions about the scalp distribution of a component, it may be necessary to include measurements from all of the electrodes.

Box 6.2 provides a few thoughts about how to describe ANOVA results in a journal article.

Interactions with Electrode Site

It is clear from figure 6.5 that the difference in P3 amplitude between the rare and frequent stimuli was larger at posterior sites than at anterior sites. This led to a significant interaction between stimulus probability and anterior-posterior electrode position, $F(2, 18) = 63.92$, $p < .001$. In addition, the probability effect for the bright stimuli was somewhat larger than the probability effect for the dim stimuli at the parietal electrodes, but there wasn't much difference at the frontal electrodes. This led to a significant

Box 6.2 The Presentation of Statistics

Inferential statistics such as ANOVAs are used to tell us how much confidence we can have in our data. As such, they are not data, but are a means of determining whether the data pattern is believable. However, many ERP results sections are written as if the inferential statistics are the primary results, with virtually no mention of the descriptive statistics or the ERP waveforms. For example, consider this excerpt from a Results section:

Results

There was a significant main effect of probability on P3 amplitude, $F(1, 9) = 4.57$*,* $p < .02$*. There was also a significant main effect of electrode site,* $F(11, 99) = 3.84$*,* $p < .01$*, and a significant interaction between probability and electrode site,* $F(11, 99) = 2.94$*,* $p < .05$ *. . .*

In this example, the reader learns that there is a significant effect of probability on the P3 wave but cannot tell whether the P3 was bigger for the probable stimuli or for the improbable stimuli. This is not an unusual example: I have seen Results sections like this many times when reviewing journal submissions. It is important to remember that inferential statistics should always be used as support for the waveforms and the means, and the statistics should never be given without a description of the actual data. And I am not alone in this view. For example, an editorial in *Perception & Psychophysics* a few years ago stated that "the point of Results sections, including the statistical analyses included there, is to make the outcome of the experiment clear to the reader. . . . Readers should be directed first to the findings, then to their analysis" (Macmillan, 1999, p. 2).

three-way interaction among probability, brightness, and anterior-posterior electrode position, $F(2, 18) = 35.17$, $p < .001$.

From this interaction, you might be tempted to conclude that different neural generator sites were involved in the responses to the bright and dim stimuli. However, as McCarthy and Wood (1985) pointed out, ANOVA interactions involving an electrode position factor are ambiguous when two conditions have different mean amplitudes. Figure 6.6 illustrates this, showing the scalp distributions that one would expect from a single generator source in two different conditions, A and B. If the magnitude of the generator's

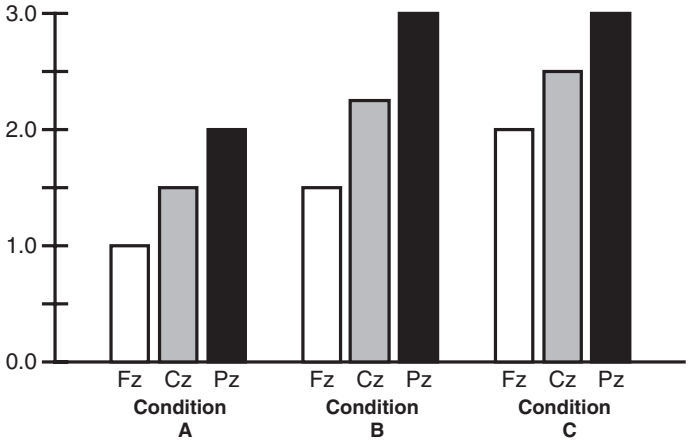


Figure 6.6 Examples of additive and multiplicative effects on ERP scalp distributions. Condition B is the same as condition A, except that the magnitude of the neural generator site has increased by 50 percent, thus increasing the voltage at each site by 50 percent. This is a multiplicative change. Condition C is the same as condition A, except that the voltage at each site has been increased by 1 μV. This is an additive change and is not what would typically be obtained by increasing the magnitude of the neural generator source.

activation is 50 percent larger in condition B than in condition A, the amplitude at each electrode will be 50 percent larger in condition B than in condition A. This is a multiplicative effect, and not an additive effect. That is, the voltage increases from 1 μV to 1.5 μV at Fz and increases from 2 μV to 3 μV at Pz, which is a 0.5 μV increase at Fz and a 1 μV increase at Pz. Condition C in figure 6.6 shows an additive effect. In this condition, the absolute voltage at each site increases by 1 μV, which is not the pattern that would result from a change in the amplitude of a single generator source. Thus, when a single generator source has a larger magnitude in one condition than in another condition, an interaction between condition and electrode site will be obtained (as in condition A versus condition B), whereas a change involving multiple generator sites may sometimes produce a purely additive effect (as in condition A versus condition C).

To determine whether an interaction between an experimental condition and electrode site really reflects a difference in the internal generator sources, McCarthy and Wood (1985) proposed *normalizing* the data to remove any differences in the overall amplitudes of the conditions. Specifically, the data are scaled so that the voltages across the electrode sites range between 0 and 1 in both conditions. To do this, you simply find the electrode sites with the largest and smallest amplitudes in each condition, and for each condition compute a new value according to the formula:

$$\text{New value} = (\text{old value} - \text{minimum value}) \div (\text{maximum value} - \text{minimum value})$$

Once these normalized values have been computed, there will be no difference in amplitude between the conditions, and the condition × electrode site interaction is no longer distorted by the multiplicative relationship between the magnitude of the internal generator and the distribution of voltage across electrodes. Thus, any significant interaction obtained with the normalized values must be due to a change in the relative distribution of internal brain activity.

But there is a problem. Many labs used this procedure for almost two decades, and then Urbach and Kutas (2002) convincingly demonstrated that it doesn't work. For a variety of subtle technical reasons, normalization procedures will fail to adjust accurately for the multiplicative interactions that arise when ANOVAs are conducted with electrode as a factor. A significant condition × electrode site interaction may be obtained after normalization even if there is no difference in the relative distribution of internal brain activity, and a real difference in the relative distribution of internal brain activity may not yield a significant condition × electrode site interaction (even with infinite statistical power). Thus, you should not use normalization procedures, and there is simply no way to determine the presence or absence of a change in the relative distribution of internal brain activity by examining condition × electrode site interactions. It just cannot be done.

There is also another problem with normalization, but it's conceptual rather than technical. Many researchers have looked at condition × electrode site interactions to determine whether or not the same brain areas are active in different experimental conditions. Even if normalization worked correctly, it would be impossible to make claims of this sort on the basis of a condition × electrode site interaction. The problem is that this interaction will be significant when exactly the same generators are active in both conditions as long as they differ in relative magnitude. That is, if areas A and B have amplitudes of six and twelve units in one condition and eight and nine units in another condition, this will lead to a change in scalp distribution that will produce a condition × electrode site interaction. Moreover, even if there is no difference in the magnitude of the generators across conditions, but there is a latency difference in one of the two generators across conditions, it is likely that the measured scalp distribution at any given time point will differ across conditions. Thus, you cannot draw strong conclusions about differences in which generator sources are active on the basis of ANOVA results.

Violation of ANOVA Assumptions

It is well known that the ANOVA approach assumes that the data are normally distributed and that the variances of the different conditions are identical. These assumptions are often violated, but ANOVA is fairly robust when the violations are mild to moderate, with very little change in the actual probability of a type I error (Keppel, 1982). Unless the violations of these assumptions are fairly extreme (e.g., greater than a factor of two), you just don't need to worry about them. However, when using within-subjects ANOVAs, another assumption is necessary, namely homogeneity of covariance (also called *sphericity*). This assumption applies only when there are at least three levels of a factor. For example, imagine an experiment in which each subject participates in three

conditions, C1, C2, and C3. In most cases, a subject who tends to have a high value in C1 will also tend to have high values in C2 and C3; in fact, this correlation between the conditions is the essential attribute of a within-subjects ANOVA. The assumption of homogeneity of covariance is simply the assumption that the degree of correlation between C1 and C2 is equal to the degree of correlation between C2 and C3 and between C1 and C3. This assumption does not apply if there are only two levels of a factor, because there is only one correlation to worry about in this case.

The homogeneity-of-covariance assumption is violated more often by ERP experiments than by most other types of experiments because data from nearby electrodes tend to be more correlated than data from distant electrodes. For example, random EEG noise at the Fz electrode will spread to Cz more than to Pz, and the correlation between the data at Fz and the data at Cz will be greater than the correlation between Fz and Pz. In addition, ANOVA is not very robust when the homogeneity of covariance assumption is violated. Violations will lead to artificially low p-values, such that you might get a p-value of less than .05 even when the actual probability of a type I error is 15 percent. Thus, it is important to address violations of this assumption.

The most common way to address this is to use the Greenhouse-Geisser epsilon adjustment (see Jennings & Wood, 1976), which counteracts the inflation of type I errors produced by heterogeneity of variance and covariance (which is also called *nonsphericity*). For each F-value in an ANOVA that has more than 1 degree of freedom in the numerator, this procedure adjusts the degrees of freedom downward—and hence the p-value upward—in a manner that reflects the degree of nonsphericity for each F-value. Fortunately, most major statistics packages provide this adjustment, and it is therefore easy to use.

We used the Greenhouse-Geisser adjustment in the statistical analysis of the P3 amplitude data discussed above. It influenced only the main effects and interactions involving the electrode factors, because the other factors had only two levels (i.e., frequent

vs. rare and bright vs. dim). For most of these F-tests, the adjustment didn't matter very much because the unadjusted effects were either not significant to begin with or were so highly significant that a moderate adjustment wasn't a problem (e.g., an unadjusted p-value of .00005 turned into an adjusted p-value of .0003). However, there were a few cases in which a previously significant p-value was no longer significant. For example, in the analysis of the normalized ANOVA, the main effect of anterior-posterior electrode site was significant before the adjustment was applied ($F(2, 18) = 4.37$, $p = .0284$) but was no longer significant after the adjustment ($p = .0586$). This may seem like it's not such a great thing, because this effect is no longer significant after the adjustment. However, the original p-value was not accurate, and the adjusted p-value is closer to the actual probability of a type I error. In addition, when very large sets of electrodes are used, the adjustments are usually much larger, and spurious results are quite likely to yield significant p-values when no adjustment is used.

In my opinion, it is absolutely necessary to use the Greenhouse-Geisser adjustment—or something comparable[2]—whenever there are more than two levels of a factor in an ANOVA, especially when one of the factors is electrode site. And this is not just my opinion, but reflects the consensus of the field. Indeed, the journal *Psychophysiology* specifically requires that authors either use an adjustment procedure or demonstrate that their data do not violate the sphericity assumption. In addition, most ERP-savvy reviewers at other journals will require an adjustment procedure. Even if you could ''get away'' with not using the adjustment, you should use it anyway, because without it your p-values may be highly distorted.

Follow-Up Comparisons

Once you find a significant effect in an ANOVA, it is often necessary to do additional analyses to figure out what the effect means. For example, if you have three experimental conditions and the

ANOVA indicates a significant difference among them, it is usually necessary to conduct additional analyses to determine which of these three conditions differs from the others. Or, if you have a significant interaction, you may need to conduct additional analyses to determine the nature of this interaction. In our P3 experiment, for example, we found an interaction between stimulus brightness and stimulus probability, with a larger effect of probability for the bright stimuli than for the dim stimuli. One could use additional analyses to ask whether the probability effect was significant only for the bright stimuli and not for the dim stimuli, or whether the brightness effect was significant only for the rare stimuli and not for the frequent stimuli.

The simplest way to answer questions such as these is to run additional ANOVAs on subsets of the data. For example, we ran an ANOVA on the data from the bright stimuli and found a significant main effect of probability; we also ran an ANOVA on the data from the dim stimuli and again found a significant main effect of probability. Thus, we can conclude that probability had a significant effect for both bright and dim stimuli even though the original ANOVA indicated that the probability effect was bigger for bright stimuli than for dim stimuli.

When a subset of the data are analyzed in a new ANOVA, it is possible to use the corresponding error term from the original ANOVA (this is called ''using the pooled error term''). The advantage to using the pooled error term is that it has more degrees of freedom, leading to greater power. The disadvantage is that any violations of the homogeneity of variance and covariance assumptions will invalidate this approach. Because ERP data almost always violate these assumptions, I would recommend against using the pooled error term in most cases.

There are several additional factors that you must consider when conducting additional analyses of this nature. For example, the more p-values you compute, the greater is the chance that one of them will be significant by chance. However, these additional factors are no different in ERP research than in other areas, and I will

not try to duplicate the discussions of these issues that appear in standard statistics texts.

Analyzing Multiple Components

In the standard approach to ERP analysis, a separate ANOVA is conducted for each peak that is measured. It would be possible to include data from separate peaks, with peak as a within-subjects ANOVA factor, but this would not be in the univariate ''spirit'' of the ANOVA approach. That is, the different components are not really measurements of the same variable under different conditions. Of course, performing a separate ANOVA for each peak can lead to a proliferation of p-values, increasing your experimentwise error, and this can be a substantial problem if you are measuring many different peaks. As chapter 2 discussed, it is usually best to focus an experiment on just one or two components rather than ''fishing'' for effects in a large set of components, and this minimizes the experimentwise error. It is also possible to enter the data from all of the peaks into a single MANOVA, but multivariate statistics are beyond the scope of this book (for more information, see Donchin & Heffley, 1978; Vasey & Thayer, 1987).

When it is necessary to analyze the data from multiple peaks, it is important to avoid the problem of temporally overlapping components as much as possible. For example, the effects of stimulus probability on the P2, N2, and P3 waves shown in figure 6.5 overlap with each other, especially at the central and parietal electrode sites. In particular, the elimination of the P2 effect at Cz and Pz for the dim stimuli could be due to an increase in P2 latency, pushing the P2 wave into the latency range of the N2 wave and leading to cancellation of the P2 wave by the N2 wave.

When you are faced with data such as these, you really have three choices. The first is to simply ignore the P2 and N2 waves and focus on the P3 wave, which is so much larger than the P2 and N2 waves that it is unlikely to be substantially distorted by them. This is a very reasonable strategy when the experiment was

designed to test a specific hypothesis about a single component and the other components are irrelevant. A second approach is to use a more sophisticated approach such as PCA or ICA. This can also be a reasonable approach, as long as you are careful to consider the assumptions that lie behind these techniques. The third approach is to use simple measurement and analysis techniques, but with measurement windows and electrode locations designed to minimize the effects of overlapping components (however, you must be cautious about the possibility that the overlapping components might still distort the results).

As an example of this third approach, we measured P2 amplitude as the mean voltage between 125 and 275 ms, which spanned most of the P2 wave for both the bright and dim stimuli. In addition, to minimize the effects of the overlapping components, we measured the P2 wave only at the frontal electrode sites, where the P2 wave was fairly large and the other components were fairly small. In this analysis, we found a significant main effect of stimulus probability, $F(1, 9) = 6.13$, $p < .05$, a significant main effect of stimulus brightness, $F(1, 9) = 22.12$, $p < .002$, and a significant probability $\times$ brightness interaction, $F(1, 9) = 5.44$, $p < .05$. Although we can't be 100 percent certain that these effects aren't distorted somewhat by the overlapping N2 wave and the effects of stimulus brightness on P2 latency, they correspond well with the waveforms shown in figure 6.5 and are probably real.

The Bottom Line

In this section, I have discussed the most common approach to performing statistical analyses on ERP data. As I mentioned at the beginning of the section, this approach is not without flaws, but it also has two important positive attributes. First and foremost, it is a conventional approach, which means that people can easily understand and evaluate results that are analyzed in this manner. Second, even though a rigid .05 criterion is somewhat silly when the assumptions of ANOVA are violated and when a large number of

p-values are calculated, it is reasonably conservative and allows everyone to apply a common standard. My ultimate recommendation, then, is to use the standard approach when presenting results to other investigators, but to rely on high levels of statistical power, replication, and common sense when deciding for yourself whether your results are real.

Suggestions for Further Reading

The following is a list of journal articles and book chapters that provide useful information about ERP measurement and analysis techniques.

Coles, M. G. H., Gratton, G., Kramer, A. F., & Miller, G. A. (1986). Principles of signal acquisition and analysis. In M. G. H. Coles, E. Donchin & S. W. Porges (Eds.), *Psychophysiology: Systems, Processes, and Applications* (pp. 183–221). New York: Guilford Press.

Donchin, E., & Heffley, E. F., III. (1978). Multivariate analysis of event-related potential data: A tutorial review. In D. Otto (Ed.), *Multidisciplinary Perspectives in Event-Related Brain Potential Research* (pp. 555–572). Washington, D.C.: U.S. Government Printing Office.

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

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Jr., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. P., & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology*, *37*, 127–152.

Rosler, F., & Manzey, D. (1981). Principal components and varimax-rotated components in event-related potential research: Some remarks on their interpretation. *Biological Psychology*, *13*, 3–26.

Ruchkin, D. S., & Wood, C. C. (1988). The measurement of event-related potentials. In T. W. Picton (Ed.), *Human Event Related Potentials* (pp. 121–137). Amsterdam: Elsevier.

Squires, K. C., & Donchin, E. (1976). Beyond averaging: The use of discriminant functions to recognize event related potentials elicited by single auditory stimuli. *Electroencephalography and Clinical Neurophysiology*, *41*, 449–459.

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

Vasey, M. W., & Thayer, J. F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology*, *24*, 479–486.

Wood, C. C., & McCarthy, G. (1984). Principal component analysis of event-related potentials: Simulation studies demonstrate misallocation of variance across components. *Electroencephalography and Clinical Neurophysiology*, *59*, 249–260.