

## Ten Simple Rules for Designing and Interpreting ERP Experiments

Steven J. Luck

University of Iowa

### Overview

This chapter discusses some of the central issues in the design and interpretation of ERP experiments. Many design and interpretation issues are unique to a given content area, but many principles apply to virtually all ERP studies and these common principles are the focus of this chapter. The chapter begins with a discussion of the nature of ERP components, an issue that lies at the heart of the design and interpretation of ERP experiments. It then discusses some general principles of design and interpretation. Finally, it discusses a simple and yet vexing issue that must be addressed by any experimental design, namely the number of trials per subject. Throughout the chapter, the most significant points are distilled into a set of ten simple rules for designing and interpreting ERP experiments.

### Peaks and Components

The term *ERP component* refers to one of the most important but most nebulous concepts in ERP research. An ERP waveform unambiguously consists of a series of peaks and troughs, but these voltage deflections reflect the sum of several relatively independent underlying or *latent* components. It is extremely difficult to isolate the latent components so that they can be measured independently, and this is the single biggest roadblock to designing and interpreting ERP experiments. Consequently, one of the keys to successful ERP research is to distinguish between the observable peaks of the waveform and the unobservable latent components. This section describes several of the factors that make it difficult to assess the latent components, along with a set of “rules” for avoiding misinterpreting the relationship between the observable peaks and the underlying components.

The relationship between the visible ERP peaks and the latent ERP components is illustrated in panels A-C of Figure 1. Panel A shows an ERP waveform, and panel B shows a set of three latent ERP components that when summed together equal the ERP waveform in panel A. When several voltages are present simultaneously in a conductor such as the brain, the combined effect of the individual voltages is exactly equal to their sum, so it is quite reasonable to think about ERP waveforms as an expression of several summed latent components. In most ERP experiments, the researchers want to know how a specific latent component is influenced by an experimental manipulation, but we don’t have direct access to the latent components and must therefore make inferences about the latent components from the observed ERP waveforms. This is usually more difficult than it might seem, and the first step is to realize that the maximum and minimum voltages (i.e., the peak amplitudes) in an observed ERP waveform are not usually a good reflection of the latent components. For example, the latency of peak 1 in the ERP waveform in Panel A is much earlier than the peak latency of compo-

nent C1 in Panel B. This leads to our first rule of ERP experimental design and interpretation:

*Rule #1- Peaks and components are not the same thing. There is nothing special about the point at which the voltage reaches a local maximum or minimum.*

Researchers often quantify ERP waveforms by measuring the amplitude and latency of the voltage peaks, either implicitly or explicitly assuming that these measures provide a good means of assessing the magnitude and timing of a particular latent component. This is not usually a good assumption, and it leads to many errors in interpretation. Strategies for avoiding this problem are discussed in a later section of this chapter.

Panel C of Figure 1 shows another set of latent components that also sum together to equal the ERP waveform shown in panel A. In this case, the relatively short duration and rounded shape of peak 2 in panel A bears little resemblance to the long duration component C2’ in panel C. This leads to our second rule:

*Rule #2- It is impossible to estimate the time course or peak latency of a latent ERP component by looking at a single ERP waveform – there may be no obvious relationship between the shape of a local part of the waveform and the underlying latent components.*

Violation of this rule is especially problematic when two or more ERP waveforms are being compared. For example, consider the ERP waveforms shown in panel D of Figure 1. The solid waveform represents the sum of the three latent components shown in panel C (and is the same ERP waveform as in panel A). The dashed waveform shows the effect of decreasing component C2’ by 50%. To make this a bit more concrete, you can think of these waveforms as the response to an attended stimulus and an unattended stimulus, respectively, such that ignoring the stimulus leads to a 50% decline in the amplitude of component C2’. Without knowing the underlying component structure, it would be tempting to conclude from the ERP waveforms shown in panel D that the attentional manipulation does not merely cause a decrease in the amplitude of component C2’ but also causes: (a) a decrease in the amplitude of component C1’, (b) an increase in the amplitude of component C3’, and (c) a decrease in the latency of component C3’. In other words, the finding of an effect that overlaps with multiple peaks in the ERP waveform tends to be interpreted as reflecting changes in multiple underlying components, but this is often not the case. Alternatively, you might conclude from the waveforms in panel D that the attentional manipulation adds an additional, long-duration component that would not otherwise be present at

all. This would also be an incorrect conclusion, which leads us to:

*Rule #3- It is extremely dangerous to compare an experimental effect (i.e., the difference between two ERP waveforms) with the raw ERP waveforms.*

This example raises an important point about the relationship between amplitude and latency. Although the amplitude and latency of a latent component are conceptually independent, amplitude and latency often become confounded when ERP waveforms are measured. Consider, for example, the relatively straightforward correspondence between the peaks in panel A of Figure 1 and the latent components in panel B of the figure. Panel E of the figure shows the effects of increasing the amplitude of the first latent component on the summed ERP activity. When the amplitude of component A is increased by 50%, this creates an increase in the latency of both peak 1 and peak 2 in the summed waveform, and it also causes a decrease in the peak amplitude of peak 2. Panel F illustrates the effect of doubling the amplitude of the component C3, which causes a decrease in the amplitude and the latency of the second peak. Once again, this shows how the peak voltage in a given time range is a poor measure of the underlying ERP components in that latency range. This leads to our next rule:

*Rule #4- Differences in peak amplitude do not necessarily correspond to differences in component size, and differences in peak latency do not necessarily correspond to changes in component timing.*

In the vast majority of ERP experiments, the ERP waveforms are isolated from the EEG by means of signal-averaging procedures. It is tempting to think of signal-averaging as a process that simply attenuates the nonspecific EEG, allowing us to see what the single-trial ERP waveforms look like. However, to the extent that the single-trial waveform varies from trial to trial, the averaged ERP may provide a distorted view of the single-trial waveforms, particularly when component latencies vary from trial to trial. This is illustrated in panels G and H of Figure 1. Panel G illustrates three single-trial ERP waveforms (without any EEG noise), with significant latency variability across trials, and panel H shows the average of those three single-trial waveforms. The averaged waveform differs from the single-trial waveforms in two significant ways. First, it is smaller in peak amplitude. Second, it is more spread out in time. In addition, even though the waveform in panel H is the average of the waveforms in panel G, the onset time of the averaged waveform in panel H reflects the onset time of the earliest single-trial waveform and not the average onset time. This leads to our next rule:

*Rule #5- Never assume that an averaged ERP waveform accurately represents the single-trial waveforms.*

Fortunately, it is often possible to measure ERPs in a way that avoids the distortions created by the signal-averaging process. For example, the area under the curve in the averaged waveform shown in panel H is equal to the

average of the area under the single-trial curves in panel G. In most cases, measurements of area amplitude (i.e., mean amplitude over a fairly broad time interval) are superior to measurements of peak amplitude. Similarly, it is possible to find the time point that divides the area into two equal halves, and this can be a better measurement of latency than peak measures (see Hansen & Hillyard, 1984; Luck, 1998).

It is worth mentioning that the five rules presented so far have been violated in a very large number of published ERP experiments. There is no point in cataloging the cases, especially given that some of my own papers would be included in the list. However, violations of these rules significantly undermine the strength of the conclusions that can be drawn from these experiments. For new students of the ERP technique, it would be worth reading a large set of ERP papers and trying to identify both violations of these rules and methods for avoiding the pitfalls that the rules address.

### **What is an ERP Component?**

So how can we accurately assess changes in latent components on the basis of the observed ERP waveforms? Ideally, we would like to be able to take an averaged ERP waveform and use some simple mathematical procedure to recover the actual waveforms corresponding to the components that sum together to create the recorded ERP waveform. We could then measure the amplitude and the latency of the isolated components, and changes in one component would not influence our measurement of the other components. Unfortunately, just as there are infinitely many generator configurations that could give rise to a given ERP scalp distribution, there are infinitely many possible sets of latent components that could be summed together to give rise to a given ERP waveform. In fact, this is the basis of Fourier analysis: Any waveform can be decomposed into the sum of a set of sine waves. Similarly, techniques such as principal components analysis (PCA) and independent components analysis (ICA) use the correlational structure of a data set to derive a set of basis components that can be added together to create the observed waveforms. Localization techniques can also be used to compute component waveforms at the site of each ERP generator source. Unfortunately, these techniques have significant limitations, as will be discussed later in this section.

All techniques for estimating the latent components are based on assumptions about what a component is. In the early days of ERP research, a component was defined primarily on the basis of its polarity, latency, and general scalp distribution. For example, the P3A and P3B components were differentiated on the basis of the earlier peak latency and more frontal distribution of the P3A component relative to the P3B component. However, polarity, latency, and scalp distribution do not really capture the essence of a component. For example, the peak latency of the P3B component may vary by hundreds of milliseconds depends on the difficulty of the target-nontarget discrimination (Johnson, 1986), and the scalp distribution of the auditory N1 wave depends on the pitch of the eliciting stimulus in a manner that corresponds with the tonotopic map of

auditory cortex (Bertrand, Perrin, & Pernier, 1991). Even polarity may vary: The C1 wave, which is generated in area V1 of visual cortex, is negative for upper-field stimuli and positive for lower-field stimuli due to the folding pattern of area V1 in the human brain (Clark, Fan, & Hillyard, 1995). Consequently, most investigators now define components in terms of a combination of computational function and neuroanatomical generator site. Consistent with this approach, my own definition of the term ERP component is *scalp-recorded neural activity that is generated in a given neuroanatomical module when a specific computational operation is performed*. By this definition, a component may occur at different times under different conditions, as long as it arises from the same module and represents the same cognitive function. The scalp distribution and polarity of a component may also vary according to this definition, because the same cognitive function may occur in different parts of a cortical module under different conditions

Techniques such as PCA and ICA use the correlational structure of an ERP data set to define a set of components, and these techniques therefore derive components that are based on functional relationships. Specifically, different time points are grouped together as part of a single component to the extent they tend to vary in a correlated manner, as would be expected for time points that reflect a common cognitive process. The PCA technique, in particular, is problematic because it does not yield a single, unique set of underlying components without additional assumptions (see, e.g., Rosler & Manzey, 1981). That is, PCA really just provides a means of determining the possible set of latent component waveshapes, but additional assumptions are necessary to decide on one set of component waveshapes (and there is typically no way to verify that the assumptions are correct). The ICA technique appears to be a much better approach, because it uses both linear and nonlinear relationships to define the components. However, any correlation-based method will have significant limitations. One limitation is that when two separate cognitive processes covary, they may be captured as part of a single component even if they occur in very different brain areas and represent different computational functions. For example, if all the target stimuli in a given experimental paradigm are transferred into working memory, an ERP component associated with target detection may always be accompanied by a component associated with working memory encoding, and this may lead PCA or ICA to group them together as a single component. Another very important limitation is that, when a component varies in latency across conditions, both PCA and ICA will treat this single component as multiple components. Thus, correlation-based techniques may sometimes be useful for identifying latent ERP components, but they do not provide a magic bullet for determining which components are influenced by an experimental manipulation.

Techniques for localizing ERPs can potentially provide measures of the time course of activity within anatomically defined regions. In fact, this aspect of ERP localization techniques might turn out to be just as important as the ability to determine the neuroanatomical locus of an ERP effect. However, there are no foolproof techniques for

localizing ERPs at present, and we may never have techniques that allow direct and accurate ERP localization. Thus, this approach to identifying latent ERP components is not generally practical at the present time.

### **Avoiding Ambiguities in Interpreting ERP Components**

The preceding sections of this chapter are rather depressing, because it seems that there is no perfect and general method for measuring latent components from observed ERP waveforms. This is a major problem, because many ERP experiments make predictions about the effects of some experimental manipulation on a given component, and the conclusions of these experiments are valid only if the observed effects really reflect changes in that component. For example, the N400 component is widely regarded as a sensitive index of the degree of mismatch between a word and a previously established semantic context, and it would be nice to use this component to determine which of two sets of words is perceived as being more incongruous. If two sets of words elicit different ERP waveforms, it is necessary to know whether this effect reflects a larger N400 for one set or a larger P3 for the other set; otherwise, it is impossible to determine whether the two sets of words differ in terms of semantic mismatch or some other variable (i.e., a variable to which the P3 wave is sensitive). Here I will describe six strategies for minimizing factors that lead to ambiguous relationships between the observed ERP waveforms and the latent components.

#### *Strategy 1: Focus on a Specific Component*

The first strategy is to focus a given experiment on only one or perhaps two ERP components, trying to keep as many other components as possible from varying across conditions. If 15 different components vary, you will have a mess, but variations in a single component are usually tractable. Of course, sometimes a “fishing expedition” is necessary when a new paradigm is being used, but don’t count on obtaining easily interpretable results in such cases.

#### *Strategy 2: Use Well-Studied Experimental Manipulations*

It is usually helpful to examine a well-characterized ERP component under conditions that are as similar as possible to conditions in which that component has previously been studied. For example, the N400 wave was discovered in a paradigm that was intended to produce a P3 wave. The fact that the experiment was so closely related to previous P3 experiments made it easy to determine that the unexpected negative wave was a new component and not a reduction in the amplitude of the P3 wave.

#### *Strategy 3: Focus on Large Components*

When possible, it is helpful to study large components such as P3 and N400. When the component of interest is very large compared to the other components, it will dominate the observed ERP waveform, and measurements of the corresponding peak in the ERP waveform will be

relatively insensitive to distortions from the other components.

#### *Strategy 4: Isolate Components with Difference Waves.*

It is often possible to isolate the component of interest by creating difference waves. For example, imagine that you are interested in assessing the N400 for two different classes of nouns, class 1 and class 2. The simple approach to this might be to present one word per second, randomly choosing words from class 1 and class 2. This would yield two ERP waveforms, one for class 1 and one for class 2, but it would be difficult to know if any differences observed between the class 1 and class 2 waveforms were due to a change in N400 amplitude or due to changes in some other ERP component. To isolate the N400, the experiment could be redesigned so that each trial contained a sequence of two words, a context word and a target word, with the target word selected from class 1 on some trials and from class 2 on others. In addition, the context and target words would sometimes be semantically related and sometimes be semantically unrelated. The N400 could then be isolated by constructing difference waves in which the ERP waveform elicited by a given word when it was preceded by a semantically related context word is subtracted from the ERP waveform elicited by that same word when preceded by a semantically unrelated context word. Separate difference waves would be constructed for class 1 targets and for class 2 targets. Because the N400 is much larger for words that are unrelated to a previously established semantic context, whereas most other ERP components are not sensitive to the degree of semantic mismatch, these difference waves would primarily reflect the N400 wave, and any differences between the class 1 and class 2 difference waves would primarily reflect differences in the N400 (for an extensive example of this approach, see Vogel, Luck, & Shapiro, 1998).

Although this approach is quite powerful, it has some limitations. First, difference waves constructed in this manner may contain more than one ERP component. For example, there may be more than one ERP component that is sensitive to the degree of semantic mismatch, so an unrelated-minus-related difference wave might consist of two or three components rather than just one. However, this is still a vast improvement over the raw ERP waveforms, which probably contain at least 10 different components. The second limitation of this approach is that it is sensitive to interactions between the variable of interest (e.g., class 1 versus class 2 nouns) and the factor that is varied to create the difference waves (e.g., semantically related versus unrelated word pairs). If, for example, the N400 amplitude is 1  $\mu$ V larger for class 1 nouns than for class 2 nouns, regardless of the degree of semantic mismatch, then the unrelated-minus-related difference waves will be identical for class 1 and class 2 nouns. Fortunately, when 2 factors influence the same ERP component, they are likely to interact multiplicatively. For example, N400 amplitude might be 20% greater for class 1 than for class 2, leading to a larger absolute difference in N400 amplitude when the words are unrelated to the context word than when they are related. Of course, the interactions

could take a more complex form that would lead to unexpected results. For example, class 1 words could elicit a larger N400 than class 2 words when the words are unrelated to the context word, but they might elicit a smaller N400 when the words are related to the context word. Thus, the use of difference waves can be very helpful in isolating specific ERP components, but care is still necessary when interpreting the results. It is also important to note that the signal-to-noise ratio of a difference wave will be lower than those of the original ERP waveforms.

#### *Strategy 5: Focus on Components that are Easily Isolated*

The previous strategy advocated using difference waves to isolate ERP components, and this strategy can be further refined by focusing on certain ERP components that are relatively easy to isolate. The best example of this is the lateralized readiness potential (LRP), which reflects movement preparation and is distinguished by its contralateral scalp distribution. Specifically, the LRP in a given hemisphere is more negative when a movement of the contralateral hand is being prepared than when a movement of the ipsilateral hand is being prepared, even if the movements are not executed. In an appropriately designed experiment, only the motor preparation will lead to lateralized ERP components, making it possible to form difference waves in which all ERPs are subtracted away except for those related to lateralized motor preparation (see Coles, 1989; Coles, Smid, Scheffers, & Otten, 1995). Similarly, the N2pc component for a given hemisphere is more negative when attention is directed to the contralateral visual field than when it is directed to the ipsilateral field, even when the evoking stimulus is bilateral. Because most of the sensory and cognitive components are not lateralized in this manner, the N2pc can be readily isolated (see, e.g., Luck, Girelli, McDermott, & Ford, 1997; Woodman & Luck, in press).

#### *Strategy 6: Component-Independent Experimental Designs*

The best strategy is to design experiments in such a manner that it does not matter which latent ERP component is responsible for the observed changes in the ERP waveforms. For example, Thorpe et al. (1996) conducted an experiment in which they asked how quickly the visual system can differentiate between different classes of objects. To answer this question, they presented subjects with two classes of photographs, pictures that contained animals and pictures that did not. They found that the ERPs elicited by these two classes of pictures were identical until approximately 150 ms, at which point the waveforms diverged. From this experiment, it is possible to infer that the brain can detect the presence of an animal in a picture by 150 ms, at least for a subset of pictures (note that the onset latency represents the trials and subjects with the earliest onsets and not necessarily the average onset time). This experimental effect occurred in the time range of the N1 component, but it may or may not have been a modulation of that component. Importantly, the conclusions of this study do not depend at all on which latent component was influenced by the experimental manipulation. Unfortunately, it is rather unusual to be able to answer a signifi-

cant question in cognitive neuroscience using ERPs in a component-independent manner, but this approach should be used whenever possible (for additional examples of this approach, see Hillyard, Hink, Schwent, & Picton, 1973; Luck, Vogel, & Shapiro, 1996; Miller & Hackley, 1992).

### **Avoiding Confounds and Misinterpretations**

The problem of assessing latent components on the basis of observed ERP waveforms is usually the most difficult aspect of the design and interpretation of ERP experiments, and this problem is particularly significant in ERP experiments. There are other significant experimental design issues that are applicable to a wide spectrum of techniques but are particularly salient in ERP experiments; these will be the focus of this section.

One of the most fundamental principles of experimentation is to make sure that a given experimental effect has only a single possible cause. One part of this principle is to avoid confounds, but a subtler part is to make sure that the experimental manipulation doesn't have secondary effects that are ultimately responsible for the effect of interest. For example, imagine that you observed that the mass of a heated beaker of water was greater than the mass of an unheated beaker. This might lead to the erroneous conclusion that hot water has a lower mass than cool water, even though the actual explanation is that some of the heated water turned to steam, which escaped through the top of the beaker. To reach the correct conclusion, it is necessary to seal the beakers so that water does not escape. Similarly, it is important to ensure that experimental manipulations in ERP experiments do not have unintended side effects that lead to an incorrect conclusion.

To explore how this sort of problem may arise in ERP experiments, imagine an experiment that examines the effects of stimulus discriminability on P3 amplitude. In this experiment, letters of the alphabet are presented foveally at a rate of 1 per second and the subject is required to press a button whenever the letter Q is presented. A Q is presented on 10% of trials and a randomly selected non-Q letter is presented on the other 90%. In addition, the letter Q never occurs twice in succession. In one set of trial blocks, the stimuli are bright and therefore easy to discriminate (the bright condition), and in another set of trial blocks the stimuli are very dim and therefore difficult to discriminate (the dim condition).

There are several potential problems with this seemingly straightforward experimental design, mainly due to the fact that the target letter (Q) differs from the nontarget letters in several ways. First, the target category occurs on 10% of trials whereas the nontarget category occurs on 90% of trials. This is one of the two intended experimental manipulations (the other being target discriminability). Second, the target and nontarget letters are physically different from each other. Not only is the target letter a different shape from the nontarget letters—and might therefore elicit a somewhat different ERP waveform—the target letter also occurs more frequently than any of the individual nontarget letters. To the extent that the visual system exhibits long-lasting and shape-specific adaptation to repeated stimuli, it is possible that the response to the letter Q will

become smaller than the response to the other letters. These physical stimulus differences probably won't have a significant effect on the P3 component, but they could potentially have a substantial effect on earlier components (for a detailed example, see Experiment 4 of Luck & Hillyard, 1994).

A third difference between the target and nontarget letters is that subjects make a response to the targets and not to the nontargets. Consequently, any ERP differences between the targets and nontargets could be contaminated by motor-related ERP activity. A fourth difference between the targets and the nontargets is that, because the target letter never occurred twice in succession, the target letter was always preceded by a nontarget letter, whereas nontarget letters could be preceded by either targets or nontargets. This is a common practice, because the P3 to the second of two targets tends to be reduced in amplitude. Eliminating target repetitions is usually a bad idea, however, because the response to a target is commonly very long-lasting and extends past the next stimulus and therefore influences the waveform recorded for the next stimulus. Thus, there may appear to be differences between the target and nontarget waveforms in the N1 or P2 latency ranges that actually reflect the offset of the P3 from the previous trial, which is present only in the nontarget waveforms under these conditions. This type of differential overlap occurs in many ERP experiments, and it can be rather subtle. For an extensive discussion of this issue, see Woldorff (1988).

A fifth difference between the targets and the nontargets arises when the data are averaged and a peak amplitude measure is used to assess the size of the P3 wave. Specifically, because there are many more nontarget trials than target trials, the signal-to-noise ratio is much better for the nontarget waveforms. The maximum amplitude of a noisy waveform will tend to be greater than the maximum amplitude of a clean waveform because the noise has not been "averaged away" as well. Consequently, a larger peak amplitude for the target waveform could be caused solely by its poorer signal-to-noise ratio even if the targets and nontargets elicited equally large responses.

The manipulation of stimulus brightness is also problematic, because this will influence several factors in addition to stimulus discriminability. First, the brighter stimuli are, well, brighter than the dim stimuli, and this may create differences in the early components that are not directly related to stimulus discriminability. Second, the task will be more difficult with the dim stimuli than with the bright stimuli. This may induce a greater state of arousal during the dim blocks than during the bright blocks, and it may also induce strategy differences that lead to a completely different set of ERP components in the two conditions. A third and related problem is that reaction times will be longer in the dim condition than in the bright condition, and any differences in the ERP waveforms between these two conditions could be due to differences in the time course of motor-related ERP activity (which overlaps with the P3 wave).

There are two main ways that problems such as these can be overcome. First, many of these problems can be avoided by designing the experiment differently. Sec-

ond, it is often possible to demonstrate that a potential confound is not actually responsible for the experimental effect; this may involve additional analyses of the data or additional experiments. As an illustration, let us consider several steps that could be taken to address the potential problems in P3 experiment described above:

1. A different letter could be used as the target for each trial block, so that across the entire set of subjects, all letters are approximately equally likely to occur as targets or nontargets. This solves the problem of having different target and nontarget shapes.
2. To avoid differential visual adaptation to the target and nontarget letters, a set of ten equiprobable letters could be used, with one serving as the target and the other nine serving as nontargets. Each letter would therefore appear on 10% of trials. If it is absolutely necessary that one physical stimulus occurs more frequently than another, it is possible to conduct a sequential analysis of the data to demonstrate that differential adaptation was not present. Specifically, trials on which a nontarget was preceded by a target can be compared with trials on which a nontarget was preceded by a nontarget. If no difference is obtained – or if any observed differences are unlike the main experimental effect – then the effects of stimulus probability are probably negligible.
3. Rather than asking the subjects to respond only to the targets, the subjects can be instructed to make one response for targets and another for nontargets. Target and nontarget RTs are likely to be different, so some differential motor activity may still be present for targets versus nontargets, but this is still far better than having subjects respond to the targets and not to the nontargets.
4. It would be a simple matter to eliminate the restriction that two targets cannot occur in immediate succession, thus avoiding the possibility of differential overlap from the preceding trial. However, if it is necessary to avoid repeating the targets, it is possible to construct an average of the nontargets that excludes trials preceded by a target. If this is done, then both the target and the nontarget waveforms will contain only trials on which the preceding trial was a nontarget.
5. There are two good ways to avoid the problem of peak amplitudes being larger when the signal-to-noise ratio is lower. First, as discussed above, the peak of an ERP waveform bears no special relationship to the corresponding latent component, so there is usually no reason to measure peak amplitude. Instead, component amplitude is quantified by measuring the mean amplitude over a predefined latency range. Mean amplitude has many advantages over peak amplitude, one of which is that it is not biased by the number of trials. If, for some reason, it is necessary to measure peak amplitude rather than mean amplitude, it is possible to avoid biased amplitude measures by creating the nontarget average from a randomly selected subset of the nontarget trials such that the target and nontarget waveforms reflect the same number of trials.
6. There is no simple way to compare the P3 elicited by bright stimuli versus dim stimuli without contributions from simple sensory differences. However, simple contributions can be ruled out by a control experiment in which the same stimuli are used but are viewed during a task that is unlikely to elicit a P3 wave (e.g., counting the total number of stimuli, regardless of the target-nontarget category). If the ERP waveforms for the bright and dim stimuli in this condition differ only in the 50-250 ms latency range, then the P3 differences observed from 300-600 ms in the main experiment cannot easily be explained by simple sensory effects and must instead reflect an interaction between sensory factors (e.g., discriminability) and cognitive factors (e.g., whatever is responsible for determining P3 amplitude).
7. The experiment should also be changed so that the bright and dim stimuli are randomly intermixed within trial blocks. In this way, the subject's state of arousal at stimulus onset will be exactly the same for the easy and difficult stimuli. This also tends to reduce the use of different strategies.
8. It is possible to use additional data analyses to test whether the different waveforms observed for the dim and bright conditions are due to differences in the timing of the concomitant motor potentials (which is plausible whenever RTs differ between 2 conditions). Specifically, if the trials are subdivided into those with fast RTs and those with slow RTs, it is possible to assess the size and scalp distribution of the motor potentials. If the difference between trials with fast and slow RTs is small compared to the main experimental effect, or if the scalp distribution of the difference is different from the scalp distribution of the main experimental effect, then this effect probably cannot be explained by differential motor potentials.

Most of these strategies are applicable in many experimental contexts, and they reflect a set of general principles that are very widely applicable. I will summarize these general principles in some additional rules:

*Rule #6- Whenever possible, avoid physical stimulus confounds by using the same physical stimuli across different psychological conditions. This includes "context" confounds, such as differences in sequential order.*

*Rule #7- When physical stimulus confounds cannot be avoided, conduct control experiments to assess their plausibility. Never assume that a small physical stimulus difference cannot explain an ERP effect (even at a long latency).*

*Rule #8- Be cautious when comparing averaged ERPs that are based on different numbers of trials.*

*Rule #9- Be cautious when the presence or timing of motor responses differs between conditions.*

*Rule #10- Whenever possible, experimental conditions should be varied within trial blocks rather than between trial blocks.*

## Number of Trials and Signal-to-Noise Ratio

One of the most basic parameters that must be set when designing an ERP experiment is the number of trials. When conventional averaging is used, the size of the signal will remain constant as more and more trials are added together, but the size of the noise will decrease. Thus, the overall signal-to-noise ratio is increased when the number of trials is increased. The number of trials needed to obtain an acceptable signal-to-noise ratio will depend on the size of the signal you are attempting to record and the noise level of the data. If you are focusing on a large component such as the P3 wave, and you expect your experimental manipulation to change the amplitude or latency by a large proportion, then you will need relatively few trials. If, however, you are focusing on a small component like the P3 wave or you expect your experimental effect to be small, then you will need a large number of trials. The noise level will also depend on the nature of the experiment and the characteristics of the subjects (e.g., young children and psychiatric patients typically have noisier signals than healthy young adults).

Experience is usually the best guide in selecting the number of trials. If you lack experience, then the literature can provide a guide (although you will want to see how clean the waveforms look in a given paper before deciding to adopt the same number of trials). Newcomers to the ERP technique usually dramatically underestimate the number of trials needed to obtain a reasonable signal-to-noise ratio.

In my own lab, the rule of thumb is that we need 30-60 trials per condition when looking at a large component like the P3 wave, 150-200 trials per condition when looking at a medium-sized component like the N2 wave, and 400-800 trials per condition when looking at a small component like the P1 wave. When recording from young children or psychiatric patients, you should try to double or triple these numbers.

It is important to realize that the relationship between the number of trials and the signal-to-noise ratio is a negatively accelerated function. To be precise, 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  $(1/\sqrt{N}) R$ . 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  $20 \mu\text{V}$  (i.e., if you could measure it without any EEG noise). If the EEG noise is  $50 \mu\text{V}$  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 2 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 4 trials (because  $\sqrt{4} = 2$ ). To quadruple the S/N

ratio from .4 to 1.6, it is necessary to average together 16 trials (because  $\sqrt{16} = 4$ ). Thus, doubling the S/N ratio requires 4 times as many trials, and quadrupling the S/N ratio requires 16 times as many trials. To get from a single-trial S/N ratio of 0.4 to a reasonable S/N ratio of 10.0 would require 625 trials. This relationship between the number of trials and the S/N 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 is why so many trials are needed in most ERP experiments.

It is also important to do whatever you can to reduce the size of the noise in the raw EEG. There are four main sources of noise. The first is EEG activity that is not elicited by the stimuli (e.g., alpha waves). This source of noise can often be reduced by making sure that the subjects are relaxed but alert. The second source is trial-to-trial variability in the actual ERP components due to variations in neural and cognitive activity; this is probably a minor source of variability in most cases, and it may be reduced by changing the task in ways that ensure trial-by-trial consistency.

The third source of noise is artifactual bioelectric activity, such as blinks, eye movements, muscle activity, and skin potentials. Blinks and eye movements can be detected and rejected during averaging, so they are not a large problem (unless a large proportion of trials is rejected). Of the remaining sources of bioelectric noise, skin potentials are probably the most significant problem. These potentials arise when the conductance of the skin changes (often due to perspiration) or the impedance of the electrode suddenly changes (often due to head movements). These can be minimized by keeping the recording environment cool and keeping electrode impedances low (high impedance amplifiers will not help reduce this type of artifact). The final source of noise is environmental electrical activity, such as line-frequency noise from video monitors and other electrical devices. This can be minimized by means of extensive shielding (e.g., video monitors can be placed inside shielded boxes). In general, it is worth spending considerable time and effort to set up the recording environment in a way that minimizes these sources of noise, because this can decrease the number of trials and/or subjects in a given experiment by 30-50%.

## References

- Bertrand, O., Perrin, F., & Pernier, J. (1991). Evidence for a tonotopic organization of the auditory cortex with auditory evoked potentials. *Acta Otolaryngologica*, *491*, 116-123.
- Clark, V. P., Fan, S., & Hillyard, S. A. (1995). Identification of early visually evoked potential generators by retinotopic and topographic analyses. *Human Brain Mapping*, *2*, 170-187.
- Coles, M. G. H. (1989). Modern mind-brain reading: Psychophysiology, physiology and cognition. *Psychophysiology*, *26*, 251-269.

- Coles, M. G. H., Smid, H., Scheffers, M. K., & Otten, L. J. (1995). Mental chronometry and the study of human information processing. In M. D. Rugg & M. G. H. Coles (Eds.), Electrophysiology of mind: Event-related brain potentials and cognition. (pp. 86-131). Oxford: Oxford University Press.
- Hansen, J. C., & Hillyard, S. A. (1984). Effects of stimulation rate and attribute cuing on event-related potentials during selective auditory attention. Psychophysiology, *21*, 394-405.
- Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. Science, *182*, 177-179.
- Johnson, R., Jr. (1986). A triarchic model of P300 amplitude. Psychophysiology, *23*, 367-384.
- Luck, S. J. (1998). Sources of dual-task interference: Evidence from human electrophysiology. Psychological Science, *9*, 223-227.
- Luck, S. J., Girelli, M., McDermott, M. T., & Ford, M. A. (1997). Bridging the gap between monkey neurophysiology and human perception: An ambiguity resolution theory of visual selective attention. Cognitive Psychology, *33*, 64-87.
- Luck, S. J., & Hillyard, S. A. (1994). Electrophysiological correlates of feature analysis during visual search. Psychophysiology, *31*, 291-308.
- Luck, S. J., Vogel, E. K., & Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. Nature, *382*, 616-618.
- Miller, J., & Hackley, S. A. (1992). Electrophysiological evidence for temporal overlap among contingent mental processes. Journal of Experimental Psychology: General, *121*, 195-209.
- 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.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. Nature, *381*, 520-522.
- Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. Journal of Experimental Psychology: Human Perception and Performance, *24*, 1656-1674.
- Woldorff, M. (1988). Adjacent response overlap during the ERP averaging process and a technique (Adjar) for its estimation and removal. Psychophysiology, *25*, 490.
- Woodman, G. F., & Luck, S. J. (in press). Serial deployment of attention during visual search. Journal of Experimental Psychology: Human Perception and Performance.



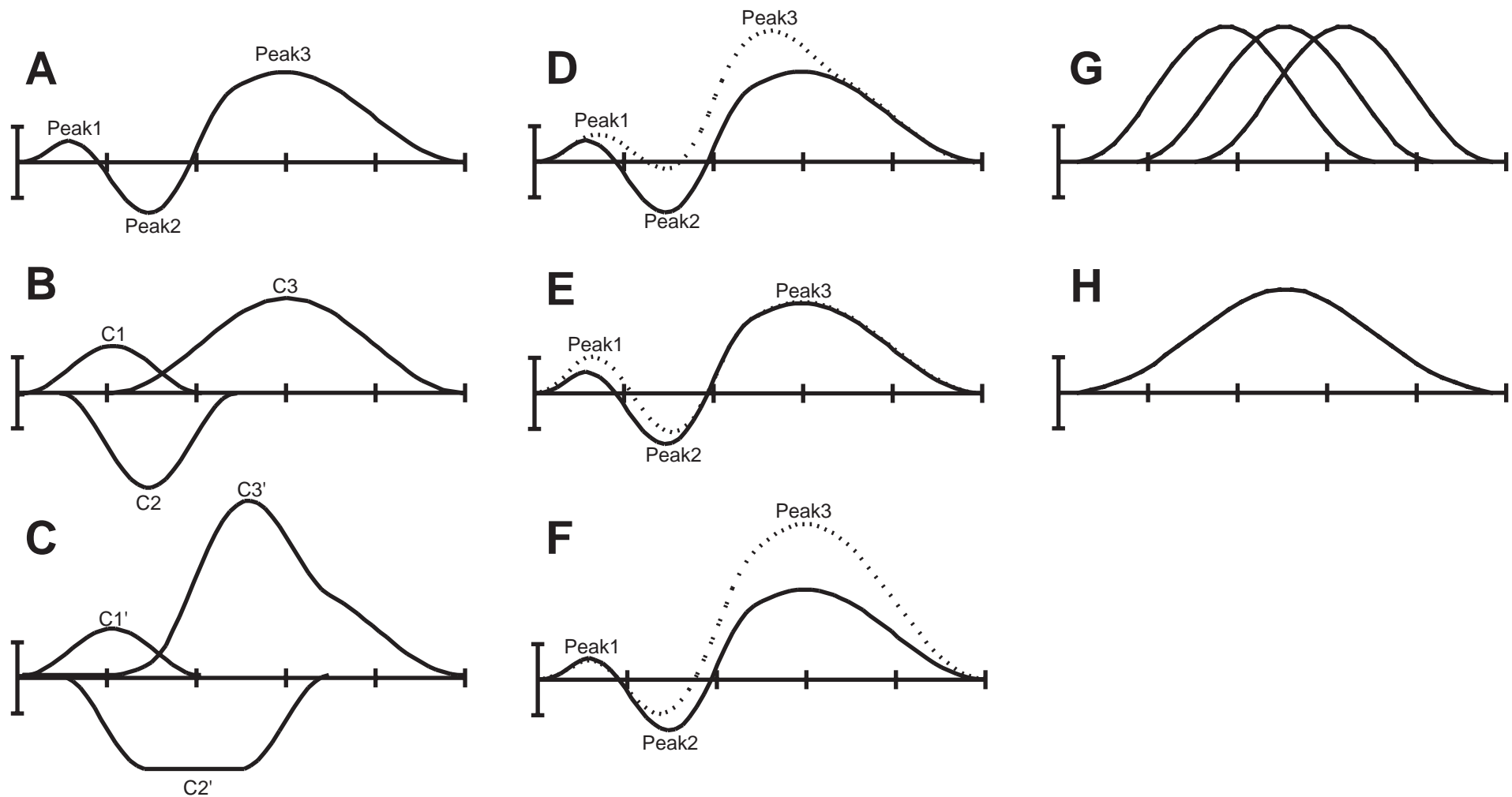


Figure 1. Examples of the latent components that may sum together to form an observed ERP waveform. Panels B and C show two different sets of latent components that could underlie the waveform shown in Panel A. Panel D shows the effect of decreasing the amplitude of component C2' by 50% (broken line) compared to the original waveform (solid line). Panel E shows how an increase in the amplitude of component C1 (broken line) relative to the original waveform (solid line) can create an apparent shift in the latencies of both peak 1 and peak 2. Panel F shows how an increase in the amplitude of component C3 (broken line) relative to the original waveform (solid line) can influence both the amplitude and the latency of peak 2. Panel G shows a component at 3 different latencies, representing trial-by-trial variations in latency; Panel H shows the average of these 3 waveforms, which is broader and has a smaller peak latency (but the same area amplitude) compared to each of the single-trial waveforms.