4 The Design of ERP Experiments

Overview

Chapter 2 described the very difficult problems that arise in trying to isolate individual ERP components. If it has been a while since you read that chapter, you might want to remind yourself by looking at figures 2.5 and 2.6. These problems can be summarized in a single paragraph:

Peaks and components are not the same thing, and this makes it difficult to isolate an underlying component from the observed ERP waveform. For example, a change in one component may influence the amplitudes and latencies of multiple peaks. Similarly, an effect that occurs in the time range that is usually associated with a given component may actually reflect a change in a different component that is also active during that time period (e.g., an effect during the time range of the N2 component may actually reflect a change in P3 amplitude during this period). Moreover, a change in component amplitude may lead to a change in peak latency, and a change in component timing may lead to a change in peak amplitude. The averaging process may also lead to incorrect conclusions because differences in latency variability will produce differences in peak amplitude. Similarly, the onset and offset of an effect in an averaged waveform will reflect the earliest single-trial onset and latest single-trial offset times rather than the average single-trial onset and offset times. Finally, it is difficult to be sure that an effect in one experiment reflects the same component that was observed in previous experiments.

In addition, chapter 3 described several examples of ERP effects that we ordinarily consider to reflect a single component but that actually reflect multiple distinct brain areas that are engaged in related (but probably not identical) processes.

Together, these problems often make it difficult to isolate highly specific psychological or neural processes in ERP experiments. This is a very real challenge because many ERP experiments make predictions about the effects of an 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 to be 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. However, the N400 and P3 components have similar

timing and scalp distributions, so it can be difficult to tell the difference between an increased N400 and a reduced P3. If one group of words produces a smaller negativity than another, this could mean that the first group of words is perceived as less incongruous (leading to a smaller N400) or is more resource-demanding (leading to a larger P3). This sort of difficulty is present in a very large proportion of ERP experiments.

But don't get depressed! You should first realize that it is always challenging to study the human mind and brain, and many of the difficulties involved in ERP research are also present in fMRI research (see the online chapter 3 supplement for a discussion of the challenges involved in linking physiological measures with the underlying processes). And the temporal resolution of the ERP technique solves many of the problems that complicate fMRI studies. Moreover, the fact that ERPs have made many important contributions to our understanding of the human mind and brain demonstrates that these problems are not insurmountable.

This chapter focuses on experimental design, with the goal of helping you figure out how to design experiments in which ERPs—despite their limitations—can provide definitive answers to important questions about the mind and brain. I will begin by describing eight time-tested strategies you can use when you design your experiments to solve or avoid the problems involved in isolating ERP components. The design phase is definitely the best time to address these potential problems; once your data have been collected, the right design will make your life much easier. After I cover those eight strategies, I will describe a variety of other experimental design problems that commonly arise in ERP experiments and provide some good solutions to these problems. An online supplement to this chapter provides several examples of experiments that successfully overcame the challenges associated with ERPs and had a real impact on our understanding of the mind and brain.

Designing experiments is my favorite part of being a scientist. The design process is where theory meets experiment, and it is a point where the inevitable imperfections of real data have not yet marred the beauty of the scientific ideas. The excellent book on the physics of EEG by Paul Nuñez and Ramesh Srinivasan notes that engineers and physicists—compared to neuroscientists and psychologists—tend to spend more of their time thinking about what experiments are worth conducting and less time implementing actual experiments (Nunez & Srinivasan, 2006). I'm not sure I agree 100% with their view of what we should be doing, but I certainly believe that people should spend more time (and intellectual effort) designing their experiments before they start collecting data (see box 4.1 for additional thoughts about experimental design).

Strategies for Avoiding Ambiguities in Interpreting ERP Components

The following eight strategies have proved to be very useful for designing experiments that minimize or avoid the problem of identifying specific ERP components. Note that they're not listed in any particular order, and you do not need to implement all strategies in all experiments. In fact, if you follow some of these strategies, others may not be necessary. Concrete examples

Box 4.1
The Craft of Experimental Design

Apple Computer Inc. is one of the world's leaders in industrial design, and one of the most influential executives at Apple is Jonathan Ive, Vice President of Design. Ive was the very first winner of the Designer of the Year Award from the Museum of Design, and he gave a very interesting interview about Apple's design process to the museum (see http://designmuseum.org/design/jonathan-ive/). In this interview, he said the following: "Perhaps the decisive factor is fanatical care beyond the obvious stuff: the obsessive attention to details that are often overlooked." I love this quote! It captures the difference between a good experimental design and a great experimental design.

I'm a fan of the Craftsman style of furniture and architecture. As shown in the following photograph, Craftsman designs take utilitarian elements such as joints and braces and turn them into aesthetic elements. By analogy, a clever counterbalancing scheme in an ERP experiment may be a thing of beauty. In addition, Craftsman designs often include elements that are hidden inside the object and cannot be seen, but these elements add to the strength and durability of the object. Similarly, if you devote "obsessive attention to details that are often overlooked," such as the temperature of the recording chamber and the quality of the signal coming from every electrode site, your results will be stronger and more durable. Craftsman designs also feature carefully selected wood that has only a few simple coats of clear finish rather than multiple layers of paint and fabric, allowing the beauty of the underlying design to be seen. I find that the most convincing ERP experiments similarly feature flawless data with only a few simple layers of processing, allowing the beauty of the underlying design to be seen.

of the application of these strategies in previous research are provided in the online supplement to this chapter.

Strategy 1: Focus on a Single Component

My first experimental design strategy is to focus a given experiment on only one or perhaps two ERP components, trying to keep all other components from varying across conditions. This reflects the tension between the conceptual and operational definitions of the term *ERP component* that were discussed in chapter 2. Manny Donchin's operational definition was that an ERP component is "a source of controlled, observable variability" (Donchin, Ritter, & McCallum, 1978, p. 354). This means that any activity that differs across a given set of conditions is equivalent to a single component. However, my conceptual definition states that an ERP component is "generated in a given neuroanatomical module when a specific computational operation is performed." If multiple processes that reflect different neuroanatomical modules and different computational operations get lumped together because they covary across the conditions of your experiment, then you will have a mess. But if you use very precise manipulations that cause only a single computational operation in a single neuroanatomical module to vary across conditions, then you will be able to isolate a single ERP component according to both the operational and conceptual definitions of *component*.

The most complicated, uninterpretable, and downright ugly results often come from experiments in which someone simply takes a previously used behavioral paradigm and runs it while recording the EEG. However, a "fishing expedition" of this sort can be very useful when you are beginning a new program of research with a task that no one has ever used with ERPs before. If you try this, you will probably find that many components vary across the conditions of the experiment and that you cannot draw any strong conclusions from the results. But this experiment may give you great ideas for more focused experiments, so it may be very worthwhile. Just make sure you treat that first experiment as a pilot study or experiment 1 in a multiple-experiment paper, and don't "pollute" the literature with complicated, uninterpretable results. For some advice about starting a new program of research, see box 4.2.

It is sometimes possible to use a factorial experimental design in which one factor is used to isolate one component and a different factor is used to isolate a different component. For example, the schizophrenia study described in chapter 1 (see figure 1.4) used a rare/frequent manipulation to isolate the P3 wave, and this was factorially crossed with a left-hand/right-hand manipulation to isolate the LRP (Luck et al., 2009). Emily Kappenman and I took this even further in a proof-of-concept study in which we used four different factors to isolate four different components. We call this the MONSTER paradigm (for *Manipulation of Orthogonal Neural Systems Together in Electrophysiological Recordings*) (Kappenman & Luck, 2011). Although this approach appears to be inconsistent with the strategy of focusing on a single component, it is actually an extension of this strategy because it is like running a sequence of four different experiments to isolate the four components.

Box 4.2
Getting Yourself a Phenomenon

My graduate school mentor was Steve Hillyard, who inherited his lab from his own graduate school mentor, Bob Galambos (shown in the photograph that follows). Dr. G (as we often called him) was still quite active after he retired. He often came to our weekly lab meetings, and I had the opportunity to work on an experiment with him. He was an amazing scientist who made fundamental contributions to neuroscience. For example, when he was a graduate student, he and fellow graduate student Donald Griffin provided the first convincing evidence that bats use echolocation to navigate. He was also the first person to recognize that glia are not just passive support cells (and this recognition essentially cost him his job at the time). You can read the details of his interesting life in his autobiography (Galambos, 1996) and in his *New York Times* obituary (http://www.nytimes.com/2010/07/16/science/16galambos.html).

Bob was always a font of wisdom. My favorite quote from him is this: "You've got to get yourself a phenomenon" (he pronounced *phenomenon* in a slightly funny way, like "pheeeenahmenahn"). This short statement basically means that you need to start a program of research with a robust experimental effect that you can reliably measure. Once you've figured out the instrumentation, experimental design, and analytic strategy that enables you to measure the effect reliably, then you can start using it to answer interesting scientific questions. You can't really answer any interesting questions about the mind or brain unless you have a "phenomenon" that provides an index of the process of interest. And unless you can figure out how to record this phenomenon in a robust and reliable manner, you will have a hard time making real progress. So, you need to find a nice phenomenon (like a new ERP component) and figure out the best ways to see that phenomenon clearly and reliably. Then you will be ready to do some real science! For examples, see the descriptions of how several ERP components were discovered at the end of chapter 3.

Strategy 2: 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 this component will be relatively insensitive to distortions from the other components. As an example, take a look at the large P3 and the small N2 in the schizophrenia study described in chapter 1 (see figure 1.4).

It is not always possible to focus on large components because sometimes a smaller component provides an index of the process that you're trying to study. However, you may be able to figure out a clever and nonobvious way to use the P3 or N400 component to answer the question you are asking.

Strategy 3: Hijack Useful Components from Other Domains

If you look at ERP experiments that have had a broad impact in cognitive psychology or cognitive neuroscience, you will find that many of them use a given ERP component that is not obviously related to the topic of the experiment. For example, the attentional blink experiment described in the online supplement to this chapter used the language-related N400 component to examine the role of attention in perceptual versus postperceptual processing (Luck, Vogel, & Shapiro, 1996; see also Vogel, Woodman, & Luck, 2005). The N400 has also been used to determine the stage of processing at which a specific variety of visual masking operates (Reiss & Hoffman, 2006). Similarly, although the LRP is related to motor preparation, it has been used to address the nature of perception without awareness (Dehaene et al., 1998) and syntax processing in language (van Turennout, Hagoort, & Brown, 1998). One of my former graduate students, Adam Niese, refers to this as *hijacking* an ERP component.

Strategy 4: 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, when Marta Kutas first started recording ERPs in language paradigms, she focused on the P3 wave and varied factors such as "surprise value" that had previously been shown to influence the P3 wave in predictable ways. Of course, when she used semantic mismatch to elicit surprise, she didn't observe the expected P3 wave but instead discovered the N400 component. However, the fact that her experiments were so closely related to previous P3 experiments made it easy to determine that the effect she observed was a new negative-going component and not a reduction in the amplitude of the P3 wave (as discussed in the final section of chapter 3).

In my own research, I have almost always included a manipulation of target probability in experiments that look at P3 (Luck, 1998b; Vogel, Luck, & Shapiro, 1998; Luck et al., 2009) and a manipulation of semantic/associative relatedness in experiments that look at N400 (Luck et al., 1996; Vogel et al., 1998; Vogel et al., 2005). Virtually everyone uses a manipulation of stimulus location to look at N2pc and CDA and a manipulation of response hand to look at the LRP, because these manipulations are intrinsic to the definition of these components.

Strategy 5: Use Difference Waves

This is probably the most important and widely applicable of all the experimental design strategies. The use of difference waves was discussed extensively in chapters 2 and 3 (see figures 2.5, 2.7, and 3.11), but here is an example that will serve as a reminder.

Imagine that you are interested in assessing the N400 for two different noun types, *count nouns* (words that refer to discrete items, such as *cup*) and *mass nouns* (words that refer to entities that are not divisible into discrete items, such as *water*). The simple approach would be to present one word at a time, with count nouns and mass nouns randomly intermixed, and have subjects do some simple semantic task (e.g., judge the pleasantness of each word). This would yield two ERP waveforms, one for count nouns and one for mass nouns. However, it would be difficult to know if any differences observed between the count noun and mass noun waveforms were due to differences in N400 amplitude or due to differences 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 a count noun target word on some trials and a mass noun target word on others. In addition, the context and target words would sometimes be semantically related and sometimes be semantically unrelated. You would then have four trial types:

Count noun, related to context word (e.g., "plate … cup")
Mass noun, related to context word (e.g., "rain … water")
Count noun, unrelated to context word (e.g., "sock … cup")
Mass noun, unrelated to context word (e.g., "garbage … water")

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 count nouns and mass nouns (unrelated minus related count nouns and unrelated minus related mass nouns). Each of these difference waves should be dominated by a large N400 component, with little or no contribution from other components (because most other components aren't sensitive to semantic mismatch). You could then see if the N400 was larger in the count noun difference wave or in the mass noun difference wave (a real application of this general approach is described in the online supplement to this chapter).

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 will probably contain at least ten different components.

A second limitation of this approach is that it is sensitive to the *interaction* between the variable of interest (e.g., count nouns versus mass nouns) and the factor that is varied to create the

difference waves (e.g., semantically related versus unrelated word pairs). Imagine, for example, that the N400 amplitude is 1 μV larger for count nouns than for mass nouns, regardless of the degree of semantic mismatch. If the N400 is 2 μV for related mass nouns and 12 μV for unrelated mass nouns, then it would be 3 μV for related count nouns and 13 μV for unrelated count nouns (i.e., 1 μV bigger than the values for the mass nouns). If we then made unrelated-minus-related difference waves, this difference would be 10 μV for the mass nouns (12 μV minus 10 μV) and would also be 10 μV for the count nouns (13 μV minus 3 μV). Fortunately, when two factors influence the same ERP component, they are likely to interact multiplicatively. Imagine, for example, that N400 amplitude is 50% greater for count nouns than for mass nouns. If the N400 is 2 μV for related mass nouns and 12 μV for unrelated mass nouns, then it would be 3 μV for related count nouns and 18 μV for unrelated count nouns (i.e., 50% bigger for the count nouns than for the mass nouns). An unrelated-minus-related difference wave would then be 10 μV for the mass nouns and 15 μV for the count nouns, so now we would be able to see the difference between count nouns and mass nouns in the related-minus-unrelated difference waves.

Of course, the interactions could take a more complex form that would lead to unexpected results. For example, count nouns could elicit a larger N400 than mass nouns 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, although difference waves can be very helpful in isolating specific ERP components, care is necessary when interpreting the results.

Strategy 6: Focus on Components That Are Easy to Isolate

To use strategy 4 and strategy 5, it is helpful to focus on those ERP components that are relatively easy to isolate by means of well-studied manipulations and difference waves. Not just any manipulation or any difference wave will do, because you want the difference wave to eliminate all components except for the one that tells you something about the process you are trying to study. For example, you could have an "easy" condition and a "difficult" condition and use these to make difficult-minus-easy difference waves, but the resulting difference waves would likely contain many different components reflecting the many different processes that might differ between difficult and easy conditions. This would not usually be very helpful.

Some components are easier to isolate than others, especially if you plan to use one manipulation to isolate the component and then factorially combine this manipulation with another manipulation designed to ask how this component varies across conditions. The best examples are the components that are defined by a contralateral-minus-ipsilateral difference wave, including the lateralized readiness potential (LRP), the N2pc component, and the contralateral delay activity (CDA; see chapter 3). For example, the LRP is defined by the difference in amplitude between the hemisphere contralateral to a response and the hemisphere ipsilateral to the response hand (reviewed by Smulders & Miller, 2012). The contra-minus-ipsi difference wave is sensitive to motor-related activity because of the contralateral organization of the motor system, but it subtracts away all other brain responses that are not contralaterally organized. It also subtracts away any brain activity prior to the time at which the brain has determined whether a left-hand

response or a right-hand response should be made for the current stimulus. Consequently, the LRP can be used to determine with high levels of certainty that the brain has begun to prepare a given response at a given moment in time (see review by Smulders & Miller, 2012). The LRP has been used in many high-impact studies, showing that the brain sometimes prepares an incorrect response even when the correct response is ultimately emitted (Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988), that partial results from one stage of processing are transmitted to the next stage (Miller & Hackley, 1992, described in the online supplement to this chapter), and that subliminal stimuli are processed all the way to response selection stages (Dehaene et al., 1998).

Similarly, the N2pc component is isolated by a contra-minus-ipsi difference wave relative to the location of an attended stimulus in a bilateral visual stimulus array (see figure 3.8 in chapter 3 and the review by Luck, 2012b). Because the overall array is bilateral, the initial sensory response is bilateral, as are the ERPs corresponding to higher-level postperceptual processes. The only processes that remain in the contra-minus-ipsi subtraction are those that are both influenced by the allocation of attention to the target and generated in contralaterally organized areas of visual cortex. Moreover, this contra-minus-ipsi difference can easily be combined with other experimental manipulations or group comparisons. For example, one can manipulate whether the attended item is associated with large or small rewards (Kiss, Driver, & Eimer, 2009) or is surrounded by nearby distractors (Luck, Girelli, McDermott, & Ford, 1997), or one can ask whether the contra-minus-ipsi difference varies as a function of age (Lorenzo-Lopez, Amenedo, & Cadaveira, 2008) or psychiatric diagnosis (Luck et al., 2006). This makes it possible to ask very precise questions about the operation of attention (see, e.g., Woodman & Luck, 1999, 2003b; Eimer & Kiss, 2008; Lien, Ruthruff, Goodin, & Remington, 2008).

Strategy 7: Use a Component to Study the Processes That Precede It

Strategy 7 is based on the idea—which was described previously in the section on P3 latency in chapter 3—that the occurrence of a difference between conditions logically entails that certain processes must have already occurred. For example, the P3 wave is larger for stimuli that belong to a rare category than for stimuli that belong to a frequent category, and the difference in amplitude cannot begin until the brain has begun to determine the category to which a stimulus belongs. Consequently, the presence of a difference between the rare and frequent categories indicates that the brain has determined whether the stimulus belongs to the rare or frequent category, and the brain must have begun to categorize the stimulus by the time the difference has exceeded 0 μV (assuming that the experiment is appropriately designed). The P3 does not itself reflect the categorization process; instead, categorization is a necessary precondition for the occurrence of a P3 probability effect.

A made-up example is shown in figure 4.1. The goal of this experiment is to measure the amount of time required to identify a digit, determine whether it is odd or even, and add it to another digit. In the experiment, subjects view a sequence of digits at the center of the monitor, with one digit every 1500 ± 100 ms (the reasons for this timing will be explained later in the

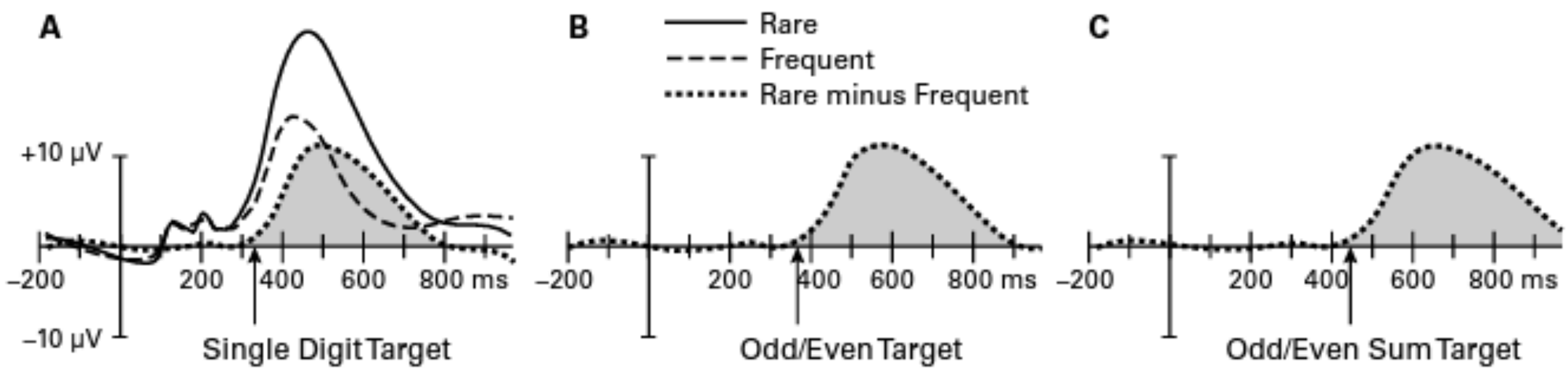


Figure 4.1
Simulated data from an imaginary oddball experiment in which the digits 0–9 are presented in the center of the screen in random order. A target category is defined at the beginning of each trial block, and the subject presses one button for target stimuli and a different button for nontarget stimuli (with the same hand). Ten percent of the stimuli are targets, and 90% are nontargets. In the single-digit condition (A), a specific digit is designated as the target, and the other nine digits serve as nontargets. ERPs elicited by stimuli in the rare (target) and frequent (nontarget) categories are shown, along with the rare-minus-frequent difference wave. In the odd/even condition (B), the target is defined as any odd digit, and the nontarget category consists of any even digit (or vice versa). Again, 10% of the stimuli are targets. Only the difference wave is shown for this condition. In the odd/even sum condition (C), the target is defined as a digit that, when added to the previous digit, is an odd number, and the nontarget is defined as a digit that, when added to the previous digit, is an even number (or vice versa). Again, 10% of the stimuli are targets. Only the difference wave is shown for this condition. For all three conditions, an arrow marks the onset latency of the difference wave, which is earliest in the single-digit condition, later in the odd/even condition, and latest in the odd/even sum condition.

chapter). Each of the 10 digits occurs in an unpredictable order. At the beginning of each trial block, a target is designated for that block. Subjects press one button when the current stimulus is a target and a different button when it is a nontarget. The target occurs on 10% of trials, and the remaining 90% are nontargets. In the *single-digit* condition, the target is defined as a specific digit (e.g., the number 3). Figure 4.1A shows the ERPs elicited by the rare stimuli (targets), the frequent stimuli (nontargets), and the rare-minus-frequent difference wave. Logically, this difference wave cannot exceed 0 μV until the brain has begun to determine whether the current stimulus is the target digit or not. Thus, the onset of the difference wave gives us an upper bound on the amount of time required for the brain to classify a number. It's an upper bound rather than the exact time because the brain might have made this classification earlier, but with no effect on the observable ERP waveform. The arrow in figure 4.1A shows the onset time of the difference, defined as the time at which it exceeds 1 μV (see chapter 9 for a discussion of methods for quantifying onset latencies).

This experiment also includes an *odd/even* condition to determine how much additional time is required to determine whether a digit is odd or even. In this condition, the target is defined as any odd digit and the nontarget is any even digit (or vice versa, counterbalanced across trial blocks). Again, the target category occurs on 10% of trials and the nontarget category occurs on 90% of trials. It presumably takes longer for the brain to determine whether a digit is odd or even than it takes to identify a specific digit, so it will take longer for the brain to determine if a given stimulus belongs to the rare category or the frequent category in this condition than in the single-digit condition. Thus, the onset of rare-minus-frequent difference should be later in

the odd/even condition than in the single-digit condition (see the arrow in figure 4.1B). Moreover, the difference in onset latency between the odd/even condition and the single-digit condition tells us how much more time is required to make an odd/even decision than is required to identify a specific digit.

The experiment also includes an *odd/even sum* condition in which the subject must determine whether the sum of the current digit and the previous digit is odd or even. Again, the sum would be odd on 10% of trials and even on 90% of trials (or vice versa). This task should take even longer than determining whether the current digit is odd or even. Thus, the onset of rare-minus-frequent difference should be even later in this condition (see the arrow in figure 4.1C) than in the odd/even condition. And the difference in onset latencies provides an estimate of the additional time required to combine two digits.

You may be asking why we included a rare/frequent manipulation in this experiment. After all, we are really interested in things like odd versus even, and combining this with rare versus frequent may seem like an unnecessary complication. However, it is in fact necessary to include the rare versus frequent manipulation. If we used 50% odd and 50% even stimuli in the odd/even condition, the ERP waveform would be virtually identical for odd stimuli and for even stimuli, and the difference wave would be a flat line. This would make it impossible to assess the time required to make the odd/even categorization. Moreover, if we compared the waveforms across the three conditions directly, without first making difference waves, then we would not be able to isolate a specific process, and we would not be able to determine the time at which the target/nontarget discrimination was made. This is a little "trick" that is often useful in ERP experiments: You take a manipulation of interest (e.g., odd versus even) and combine it with another manipulation (e.g., rare versus frequent) that will allow the manipulation of interest to generate differential ERP activity. However, you must make sure that you counterbalance the combinations so that you don't have a confound. In our odd/even condition, for example, we wouldn't want odd to be rare and even to be frequent in all trial blocks because this would confound the odd/even and rare/frequent manipulations. Instead, we would have even be rare and odd be frequent in half of the trial blocks for the odd/even condition.

Keep in mind that the P3 wave does not itself reflect the *process* of determining whether a digit is a particular number, whether it is an odd or even number, or whether the sum of two digits is odd or even. Instead, it reflects the *consequences* of making the relevant categorization. In other words, these processes must logically occur before the brain can determine whether a given stimulus falls into the rare or frequent category in this experiment. This is what I mean when I recommend using a component to assess the processes that must precede it.

Example 3 in chapter 1 describes an experiment that used this logic to study the time course of perception and categorization in schizophrenia. Subjects performed a task in which they determined whether a given stimulus was a letter or a digit; one category was rare and the other was frequent. The timing of the P3 wave, measured from the rare-minus-frequent difference wave, was virtually identical in patients and control subjects (see figure 1.4 in chapter 1 and Luck et al., 2009). This demonstrates that the patients were able to perceive and categorize

simple alphanumeric stimulus just as rapidly as the control subjects, even though behavioral RTs were substantially longer in the patients.

In these examples, ERPs are being used to measure the time course of processing, which is what ERPs do best. However, this general approach can also be used to determine whether a specific process happened at all. As described in the online supplement to this chapter, for example, the N400 component can be used to determine whether a word has been identified. Specifically, if the N400 is larger for a word that mismatches a semantic context than for a word that matches the context, the word must have been identified. This logic was used to demonstrate that words can be identified even when they cannot be consciously reported in the *attentional blink* paradigm (Luck et al., 1996; Vogel et al., 1998). The N400 does not itself reflect word identification, but word identification is a necessary precondition for seeing a difference between semantically matching and semantically mismatching words.

Strategy 8: Component-Independent Experimental Designs

Many of the previous strategies focused on ways to isolate specific ERP components. In many cases, an even better strategy is to completely sidestep the issue of identifying a specific component. For example, Thorpe, Fize, and Marlot (1996) conducted an experiment that asked how quickly the visual system can differentiate between different abstract classes of objects. To answer this question, they presented subjects with two sets 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 result, it is possible to infer that the brain can detect the presence of an animal in a picture by 150 ms (but note that the onset latency represents the trials and subjects with the earliest onsets and not necessarily the average onset time, and this is an upper bound on the initial point at which the brain detected the animal).

This experimental effect occurred in the time range of the N1 component, but it did not matter at all whether the effect consisted of a change in the amplitude of that particular component; the conclusions depended solely on the time at which the effect occurred. This was a component-independent design because the conclusions did not depend on which component was influenced by the experimental manipulation (see also the first experiment described in the online supplement to this chapter, which determined the latency at which attention influences the ERP response to a stimulus).

The use of a component to assess the processes that logically precede it (strategy 7) usually leads to a component-independent experimental design. The fact that the rare-minus-frequent difference waves were nearly identical for people with schizophrenia and healthy control subjects (see figure 1.4 in chapter 1) tells us that these two groups were able to perceive and categorize the stimuli at the same rate, even if we don't assume that the difference waves consisted of a modulation of the P3 wave. Similarly, the conclusions of the N400 experiment described in the online supplement do not actually depend on whether the effects consisted of an N400 or some other component. This is the essence of a component-independent design.

Additional examples of prior research using component-independent designs are provided in the online supplement to this chapter. For other examples of high-impact studies using this strategy (resulting in papers published in *Science* and *Nature*), see van Turennout et al. (1998) and Dehaene et al. (1998).

Common Design Problems and Solutions

Although isolating specific components is typically the most difficult aspect of ERP research, there are several other challenges that you are likely to encounter when designing your experiments. In this section, I will describe several of the most common confounds and misinterpretations in ERP research. For each problem, I will also describe one or more solutions that you can use when you design your own experiments.

Online chapter 15 will formalize this set of common problems into a list of things that you should look for when you are about to submit a paper for publication or when you are reviewing a manuscript that someone else has submitted. I'm hoping that this checklist will help expunge these common errors from the literature. See box 4.3 for a discussion of some big-picture issues in experimental design.

To make these challenges and solutions concrete, I will describe a *Gedankenexperiment* (thought experiment) that is a conglomeration of many of the bad ERP experiments I have encountered over the years (including some of my own experiments that I wish I had designed differently). This Gedankenexperiment is designed to examine the effects of task difficulty on P3 amplitude. As shown at the top of figure 4.2, the target is the letter X, and the nontarget is the letter O. The stimuli are presented at the center of the video display with a duration of 500 ms, and one letter is presented every 1000 ms (SOA = 1000 ms; ISI = 500 ms). X is presented on 20% of trials, and O is presented on the other 80%. The letter X never occurs twice in succession because the P3 is reduced for the second of two consecutive targets. Subjects press a button with the index finger of the right hand whenever an X is detected, making no response for O. In the *bright condition*, the stimuli are bright and therefore easy to discriminate; in the *dim condition*, the stimuli are very dim and therefore difficult to discriminate. The bright and dim conditions are tested in separate blocks of trials (in counterbalanced order). At the end of the experiment, typical artifact rejection and averaging procedures are conducted, and P3 amplitude is quantified as the peak voltage at the Pz electrode site, separately for the rare and frequent trials in the dim and bright conditions.

When I teach about experimental design in the ERP Boot Camp, I describe this same Gedankenexperiment, and then I ask the participants to tell me the problems they see with the design and to propose solutions for these problems. As a group, they come up with almost all of the problems that I had in mind when I designed the experiment, and they generate some interesting solutions. I would recommend that before you read further, you make a list of all the design problems that you see with this experiment and potential solutions for these problems. You can then compare your list with mine.

Box 4.3
Confounds and Side Effects

Anyone who has taken a basic course on experimental design knows that the most fundamental principle of experimentation is to make sure that a given experimental effect has only a single possible cause. This is usually discussed in terms of avoiding *confounds*. A confound occurs when more than a single factor differs between conditions. For example, if the rare stimuli in an oddball paradigm are red and the frequent stimuli are blue, then these stimuli differ in two ways (i.e., rare versus frequent is confounded with red versus blue). A true confound can almost always be detected by a careful reading of the methods section of a journal article.

A related but subtler problem occurs when the experimenter varies only one factor, but this factor has *side effects* that are ultimately responsible for the effect of interest. To take a simple example, imagine you observe that heating a beaker of water causes a decrease in the mass of the water. This might lead you 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 beaker so that water does not escape. Similarly, imagine that the P3 wave is smaller when you increase the temperature inside the recording chamber. You might conclude that warmer brains produce smaller P3 waves, but this might instead be a side effect of the fact that heating the chamber causes the subjects to become sleepier. You only varied one factor explicitly (the temperature of the chamber), but this change had multiple consequences (warmer brains, sleepier subjects, and, as described in chapter 5, more skin potentials). Side effects are more difficult to detect than confounds because they are secondary consequences of the manipulation of a single factor. Some imagination on your part may be required to realize that an experimental manipulation had a problematic side effect.

People who conduct experimental research sometimes have a condescending attitude toward people who conduct correlational research, because correlations can always be explained by some unknown third factor. But experimental effects can always be explained by some unforeseen side effect. That is, even though experiments have the advantage that one can be certain that *something* related to the experimental manipulation caused the effect, the exact cause is never known with certainty. So we should keep in mind that science is always prone to alternative explanations for both experimental and correlational research.

Sensory Confounds

There are some obvious sensory confounds in this Gedankenexperiment. First, the target is the letter X and the nontarget is the letter O, so the target and nontarget stimuli differ in terms of both shape and probability. This sort of confound is present in the vast majority of oddball ERP experiments. You might be asking why people design experiments with an obvious confound and why reviewers allow these experiments to be published. If you ask the experimenter about this kind of confound, you'll almost certainly get a response like this: "I can't imagine how that small sensory difference could produce a difference in the ERP at 400 ms, where the P3 is being measured" (see box 4.4 for a general discussion of this type of logic). This is probably true for the P3 wave, and this kind of sensory confound is therefore relatively benign for experiments focusing on late components. However, sensory confounds may produce significant effects as

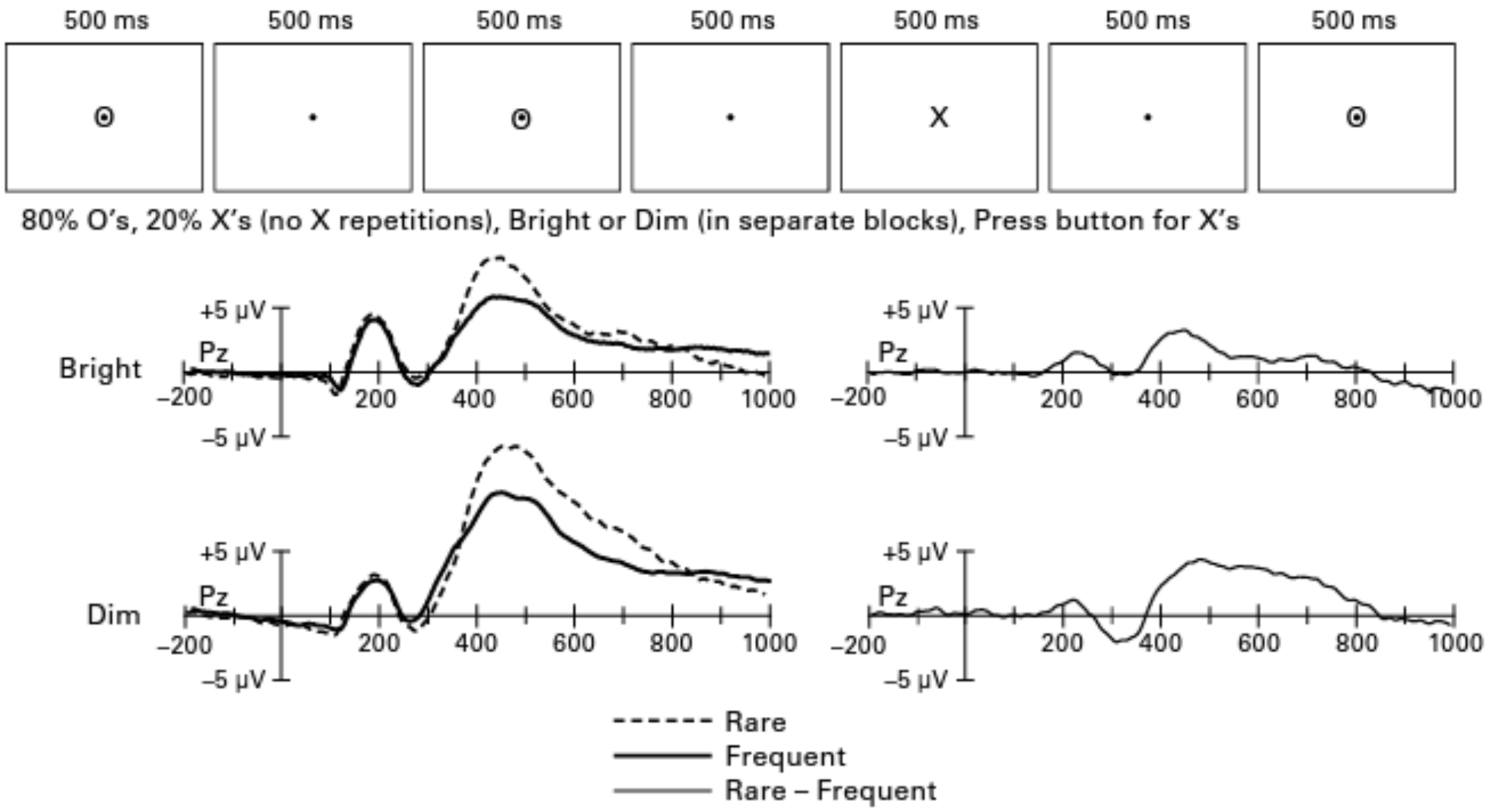


Figure 4.2
Experimental design (top) and simulated data (bottom) from the Gedankenexperiment. ERPs are overlaid for the rare and frequent stimulus categories, separately for bright stimuli and dim stimuli (left column). Rare-minus-frequent difference waves are also shown for the bright and dim stimuli (right column).

Box 4.4
Ignorance and Lack of Imagination

When someone says, "I can't imagine how that little confound could explain my results," this is a case of a general logical fallacy that philosophers call the *argument from ignorance*. In fact, it's a special case that is called (with a touch of humor) the *argument from lack of imagination*. The fact that someone can't imagine how a confound could produce a particular effect might just mean that the person doesn't have a very good imagination! I myself have occasionally used the "I can't imagine how …" type of reasoning and then found that I was suffering from a lack of imagination (see, e.g., box 4.5). But now that I realize that this is not a compelling form of argument, I usually catch myself before I say it.

late as 200–300 ms. Moreover, a benign confound is a little bit like a benign tumor; wouldn't you rather not have it? It's usually a trivial matter to design an experiment to avoid confounds of this nature, so I would recommend not marring your beautiful experimental design with an ugly little confound.

The obvious solution to this problem is to counterbalance the Xs and Os. That is, you could have rare Xs and frequent Os in half of the trial blocks and then switch to rare Os and frequent Xs in the other half. If you are working with a cognitively impaired population, this might cause some confusion, so you could instead counterbalance across subjects.

However, this experiment contains a subtler sensory confound that cannot be solved by simple counterbalancing. Specifically, a difference in the probability of occurrence between two stimuli creates differences in sensory adaptation, which will in turn create differences in the sensory response to the two stimuli. The basic idea is that when the visual system encounters a particular stimulus many times, the neurons that code that stimulus will produce smaller and smaller responses. This is known as *stimulus-specific adaptation* or *refractoriness*. The fact that O occurs more frequently than X in this Gedankenexperiment means that the neurons that code O will be more adapted than the neurons that code X, and this may lead to a smaller sensory response for the nontarget O stimuli than for the target X stimuli. This will be true even if you switch the stimuli, making O rare and X frequent in half of the trial blocks. In each block, the neurons coding the frequent stimulus will become more adapted than the neurons coding the rare stimulus, leading to a smaller sensory response for the frequent stimulus. Box 4.5 provides an example of how this sort of adaptation created a significant and replicable but bogus effect in some of my own experiments.

How might we solve this adaptation problem? The solution to this and virtually all sensory confounds is to follow the following precept:

The Hillyard principle To avoid sensory confounds, you must compare ERPs elicited by *exactly* the same physical stimuli, varying only the psychological conditions.

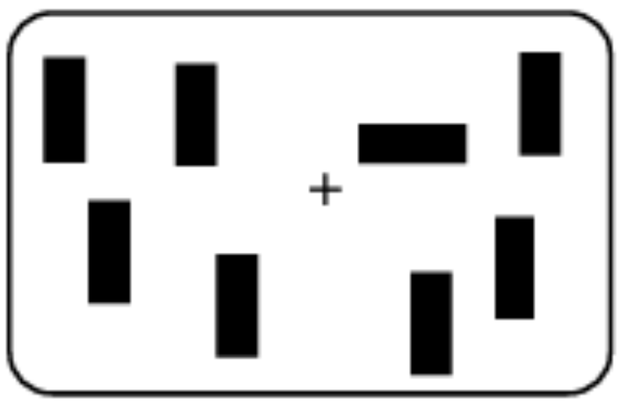
I call this *the Hillyard principle* because Steve Hillyard made his mark on the field by carefully designing his experiments to rule out confounds that had plagued other experiments (and also because this principle was continually drilled into my head when I was a grad student in the Hillyard lab). The key to this principle is that you should be able to conduct your experiments by presenting exactly the same stimulus sequences, using instructions to create the different experimental conditions. Imagine, for example, that we try to solve the sensory confound in our Gedankenexperiment by simply counterbalancing whether X or O is the rare stimulus. We can't do this simply by using the same stimulus sequences and varying whether we tell the subjects to respond to the X or to the O, because this does not change which physical stimulus is rare. Note that it is not enough to equate the ERP-eliciting stimulus across conditions; the whole sequence of stimuli must be equated to avoid all possible sensory confounds.

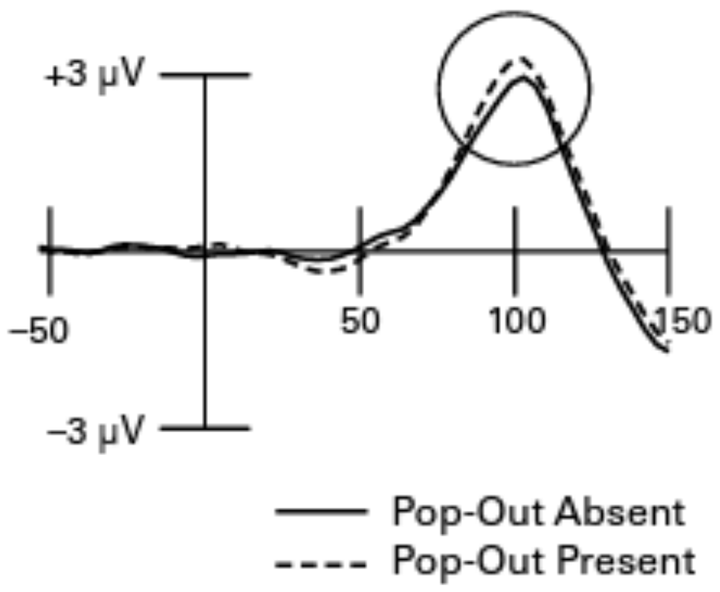
Here is how we could change the experiment to follow the Hillyard principle. Instead of using X and O as the stimuli, we could use the letters A, B, C, D, and E. Each of these letters would

Box 4.5
Example of an Adaptation Confound

Many years ago, I conducted a series of experiments in which I examined the ERPs elicited by visual search arrays consisting of seven randomly positioned "distractor" bars of one orientation and one randomly positioned "pop-out" bar of a different orientation (see the illustration in this box). In several experiments, I noticed that the P1 wave was slightly but significantly larger over the hemisphere contralateral to the pop-out item relative to the ipsilateral hemisphere (this tiny effect is highlighted with a circle in the figure). I thought this might reflect an automatic capture of attention by the pop-out item, although this didn't fit very well with what we knew about the time course of attention. My officemate, Marty Woldorff (who is now a professor at Duke), suggested that this effect might actually reflect an adaptation effect. Specifically, the location of the pop-out bar on one trial typically contained an opposite-orientation distractor bar on the previous trial, whereas the location of a given distractor bar on one trial typically contained a same-orientation distractor bar on the previous trial. Thus, the neurons that coded the distractor bars had usually responded to a bar of the same orientation in a similar location in the previous trial, whereas the neurons that coded the pop-out bar had not usually responded to a bar of that orientation at that location in the previous trial. Consequently, the neurons coding the pop-out bar may have been less adapted, generating a larger response, and this response would have been especially visible over the hemisphere contralateral to the pop-out bar.

At first, I refused to believe that this kind of adaptation could impact the P1 wave (in fact, Marty tells me that my initial response involved a phrase that is not suitable for an academic book). I couldn't imagine how this small adaptation effect could have a significant impact, especially because the screen was blank for 750 ms between trials. However, Marty kept bugging me about it, and eventually I designed an experiment to prove he was wrong. But it turned out that he was absolutely right (see experiment 4 of Luck & Hillyard, 1994b). I guess Marty's imagination was better than mine.





occur on 20% of trials. We would then have five different trial blocks, with a different letter designated as the target in each block (e.g., in one block, we would say that D is the target and all of the other letters were nontargets). We could then present exactly the same sequence of stimuli in each trial block, with the target category (e.g., D) occurring on 20% of trials and the nontarget category (e.g., A, B, C, and E) occurring on the remaining 80%. This would solve the sensory adaptation problem, because the probability of any given physical stimulus is 20%, whether it is the target stimulus or a nontarget stimulus (for an even fancier example of the experimental "backflips" that are sometimes necessary to avoid sensory confounds and satisfy the Hillyard principle, see Sawaki & Luck, 2011).

There is one small proviso to the Hillyard principle: The stimulus sequences for the different conditions should be *equivalent in principle*, but not actually the same sequences. You wouldn't want to create a situation in which people could potentially learn (whether implicitly or explicitly) the stimulus sequences by repeating them in different conditions. Nonetheless, when you generate the sequences, you should be able to use a given sequence in any condition just by changing the instructions.

Our Gedankenexperiment also contains another obvious sensory confound; namely, that the stimuli in the dim and bright conditions are physically different. The goal of the dim/bright manipulation is to change the difficulty of the task, but the dim and bright stimuli will elicit different ERP waveforms irrespective of any differences in task difficulty. For example, they would elicit different ERPs even in a passive viewing task. There are two general approaches to solving this problem. The first is to use the same stimuli for both the easy and difficult tasks and have subjects discriminate different aspects of the stimuli for the different tasks. For example, subjects could discriminate between X and O in the easy condition, and they could make a subtle size discrimination for the same stimuli in the difficult condition. To implement this, you would have four stimuli: a larger X, a smaller X, a larger O, and a smaller O. All four stimuli would be presented in every trial block, but subjects would make an X/O discrimination (ignoring size) in the easy condition and a large/small discrimination (ignoring shape) in the difficult condition. There are many different stimuli and tasks that you could use with this approach, but in all cases you would have the same physical stimuli in both the easy and difficult conditions, and difficulty would be manipulated by means of the task instructions that you give at the beginning of each trial block.

The other general approach is to use difference waves to factor out the sensory confound. To make this clear, let's imagine that we first dealt with the X/O sensory confound by using five different letters (A–E), each occurring on 20% of trials, and we instructed subjects that one of these letters was the target for a given trial block. Again, we would have bright stimuli in some trial blocks for the easy condition and dim stimuli in other trial blocks for the difficult condition. To eliminate the brightness confound, we would compute rare-minus-frequent difference waves separately for the dim and bright stimuli. First we would average across each bright letter when that letter was the target, creating a bright-rare waveform (see figure 4.2). We would also average across each bright letter when that letter was the nontarget, creating a bright-frequent

waveform. Then we would make a difference wave between these two waveforms, giving us a rare-minus-frequent difference wave for the bright stimuli. Because the bright-rare and bright-frequent waveforms were created from exactly the same stimuli, differing only in the task instruction, the difference between these waveforms no longer contains any pure sensory activity. You can see this in the difference waves shown in figure 4.2, in which the difference does not begin to deviate from 0 μV until approximately 175 ms, whereas the "parent" waveforms have a negative-going dip at 125 ms followed by a large positive wave at around 200 ms. These initial sensory responses are eliminated in the difference wave, leaving only brain activity that reflects the task-induced differential processing of the rare and frequent stimulus categories. This difference wave can then be compared with the rare-minus-frequent difference wave for the dim stimuli. Any differences between these difference waves cannot be attributed to pure sensory differences between the bright and dim stimuli and instead reflect the interaction between brightness and the task (which is what we are interested in studying in this experiment).

It is not always feasible to implement the Hillyard principle, especially when you are using naturally occurring stimulus classes. For example, a language experiment might be designed to examine the ERPs elicited by closed-class words (*the*, *for*, *with*, etc.) and open-class words (nouns, verbs, etc.), and these are by definition different stimuli. However, it is almost always possible to include a control condition that can demonstrate that the differences observed in the main condition are not a result of the sensory confound. For example, you could present closed-class and open-class words in two conditions, a main condition in which subjects are reading sentences containing these words and a control condition in which subjects are monitoring for a word that is presented in a different color. If the differences between open- and closed-class words disappear in the control condition, then you know that they are not pure sensory effects. But what if the effects are automatic and are therefore present even during the control condition? A simple approach would be to flip the words upside down and show that this causes the effects to disappear in the control condition. This would rule out low-level sensory differences as the cause of the ERP effects. Alternatively, if you are really ambitious and want to impress everyone with your experimental design abilities, you could test two groups of subjects who speak different languages, presenting open- and closed-class words of both languages to both groups of subjects. Any differences in the ERPs elicited by the open- and closed-class words that are linguistic rather than sensory in nature should be present only for subjects who speak the language in which the words are presented, and this would give you a beautiful double dissociation. It would be difficult and time-consuming for you to conduct the experiment this way, but if an experiment is worth doing, isn't it worth doing well?

I often violated the Hillyard principle early in my research career (before I started teaching other people about this principle, which has made me much more careful). However, every time I violated the Hillyard principle, I later regretted it. And I often ended up running a new experiment to rule out sensory confounds, so I would have saved a lot of time by designing the experiment properly to begin with.

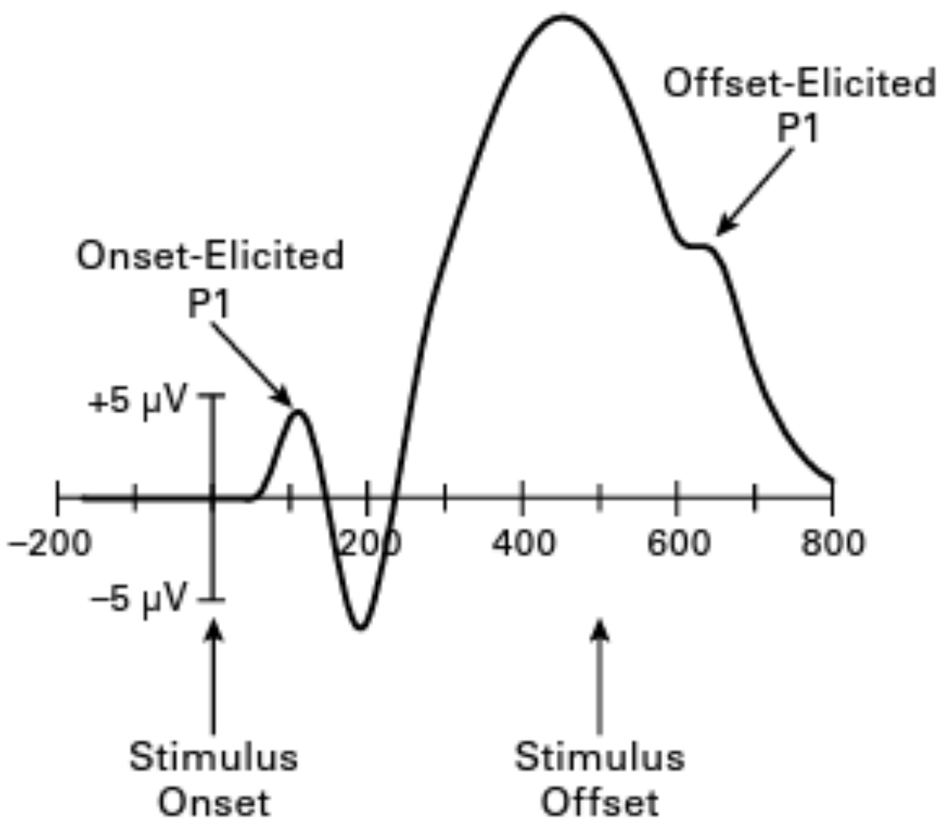


Figure 4.3
Simulated effect of a 500-ms stimulus duration. A P1 wave is triggered by the onset of the stimulus, peaking approximately 100 ms after stimulus onset, and another P1 wave is triggered by the offset of this stimulus, peaking approximately 600 ms after stimulus onset (100 ms after stimulus offset). The offset-elicited P1 adds a "bump" to the P3 wave, which might be misinterpreted as a part of the P3 wave rather than being a sensory response.

Stimulus Duration Problems

The stimuli in the Gedankenexperiment are shown for a duration of 500 ms. This duration does not create a confound, exactly, but the offset of the stimuli will generate a sensory response during the time period of the P3 wave, making the ERPs look a little weird. As shown in figure 4.3, the offset of the stimulus leads to a P1 wave approximately 100 ms after stimulus offset (600 ms after stimulus onset), which is added onto the onset-elicited P3 wave. This makes the waveforms look strange, and it could complicate the interpretation of the results because the time period of the P3 wave contains sensory activity as well as postperceptual activity.

In most cases, the offset response will be negligible if stimulus duration is short, but it becomes progressively larger as the duration becomes longer. Thus, you should choose a stimulus duration that is either so short that no significant offset activity is present or so long that the offset is after the period of time that you will be examining in your ERPs. For example, you might use a duration of 750 ms if you are interested in the ERP components that occur within the first 500 ms after stimulus onset.

How short is short enough? In the visual modality, the retina integrates photons over a period of approximately 100 ms, and a stimulus with a duration of less than ~100 ms is perceptually equivalent to a 100-ms stimulus of lower brightness (I am simplifying a bit here; if you want to know more, find a textbook that describes *Bloch's law*). Consequently, there isn't usually a reason to use a duration of less than 100 ms—you might as well just use a dimmer stimulus and a 100-ms duration. A stimulus of ~100 ms or less doesn't have a distinct onset and offset, but is just perceived as a flash. However, once the duration of the stimulus exceeds ~100 ms, increases

in duration are perceived as increases in duration rather than as increases in brightness. At this point, you will begin to perceive distinct onsets and offsets, and the ERP waveform will contain an offset response as well as an onset response. Consequently, I typically use a duration of 100 ms for visual stimuli (unless I want to use a very long duration). However, the offset response is very small for durations that are only a little more than 100 ms, and I sometimes use a duration of 200 ms if I want to give the subject a little more time to perceive the stimuli (especially when working with subjects who may have diminished sensory or cognitive abilities).

For simple auditory stimuli (e.g., sine wave tones), a duration of 50–100 ms is usually appropriate. If a sine wave (or other repeating wave) begins and ends suddenly, a clicking sound will be audible at the onset and offset. This clicking sound is great for producing very early components, but it can be distracting and can interfere with the perception of the tone's pitch (especially for short stimuli). To avoid this, the amplitude of the sound wave should ramp up over a period of 5–20 ms at the beginning of the sound and ramp down over a period of 5–20 ms at the offset of the sound (these are termed the *rise time* and *fall time* of the stimuli).

Motor Confounds

Our Gedankenexperiment also contains an obvious motor confound, because subjects make a motor 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. This is a very common confound in oddball experiments, but not as common in more sophisticated designs. In oddball experiments, one common solution is to have subjects silently count the oddballs. This doesn't really solve the problem, because the act of silently counting the targets involves additional brain activity that is not present for the standards. Also, it is difficult to assess how well the subject is performing the task (box 4.6). The best solution is usually to require the subject to press one button for the target stimuli and a different button for the standard stimuli.

Overlap Confounds

One of the most common problems that I encounter when reading ERP studies arises when the ERP elicited by one stimulus is still ongoing when the next stimulus is presented. If the overlapping activity differs across conditions, then an effect that is attributed to the processing of the current stimulus may actually be the result of continued processing of the preceding stimulus. In our Gedankenexperiment, this problem arises because the stimulus sequences were constrained so that the target letter never occurred twice in succession. Consequently, 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. Using nonrandom sequences like this is usually a bad idea, however, because the response to a target is commonly very long-lasting and can extend past the next stimulus and influence the waveform recorded for the next stimulus.

Figure 4.4 illustrates how overlap could contaminate our Gedankenexperiment (for a detailed discussion of overlap, see Woldorff, 1993). Panel A shows what the ERPs for the targets and

Box 4.6
The Problem with Counting

When I was a college student and just starting to learn about ERPs, I happened to be sitting on an airplane next to an engineer from Nicolet Instruments, a manufacturer of clinical EEG/ERP systems. The engineer told me that they were just finishing a new portable ERP system, and he asked me if I'd like to earn a little money by serving as a test subject. I was interested in both ERPs and money, so I agreed.

One of the tasks they were testing was an auditory oddball task. I was asked to sit in a dimly lit room, stare at a point on the wall, and count high-pitched target tones that were occasionally embedded among frequent low-pitched tones. I did this for about 10 minutes, but it was incredibly boring and seemed to take hours. I found it surprisingly difficult to keep track of how many high-pitched tones I had heard. I rehearsed the current count in my head until I heard the next target: "33, 33, 33, 33, 33, 33, 33, 33, 33, 33, 34, 34, 34, 34, 34, 35, 35, 35, 35 …." But as I was doing this, my mind was wandering a bit, and at some point I couldn't remember whether I was at 37 or 47. The typical procedure in counting tasks is to ask the subject how many targets he or she counted at the end of the trial block, and this experience made it clear to me that someone could detect every target and still report a very wrong number at the end. It's also possible that someone could miss one target and also misperceive one of the nontargets as a target and yet report the correct number at the end. Because of this experience, I've never been fond of counting tasks.

standards would look like without any overlap. Panel B shows the overlap that would occur with a relatively short period between trials (an SOA of approximately 700 ms). As you can see from the figure, the tail end of the P3 from the previous trial is present during the baseline period of the current trial. When the current trial is a target, the previous trial was always a nontarget, and so the overlap is simply the tail end of the ERP elicited by a nontarget (this is labeled *Overlap Y* in the figure). When the current trial is a standard, the previous trial was 80% likely to be a standard and 20% likely to be a target, and so the overlap is a weighted sum of the tail end of the standard-elicited ERP and the tail end of the target-elicited ERP (this is labeled *Overlap Z* in the figure).

In this example, the overlap from the previous trial is over by 100 ms after the onset of the current stimulus, so you might think that it wouldn't have much effect on the P3 wave on the current trial. However, the overlap is still a major problem because of *baseline correction*. For reasons that will be described in chapter 8, it is almost always necessary to use the average voltage during the prestimulus interval as a baseline, which is subtracted from the entire waveform. If the baseline is contaminated by overlap from the previous trial, the baseline correction procedure will "push" the whole waveform downward (if the contamination is from a positive component like P3) or upward (if the contamination is from a negative component). The overlap in our Gedankenexperiment is a positive voltage (from the P3 wave), so it causes the waveform to be pushed downward (which you can see in panel C of figure 4.4). The overlap is greater when the current stimulus is a standard than when it is a target (because the preceding stimulus

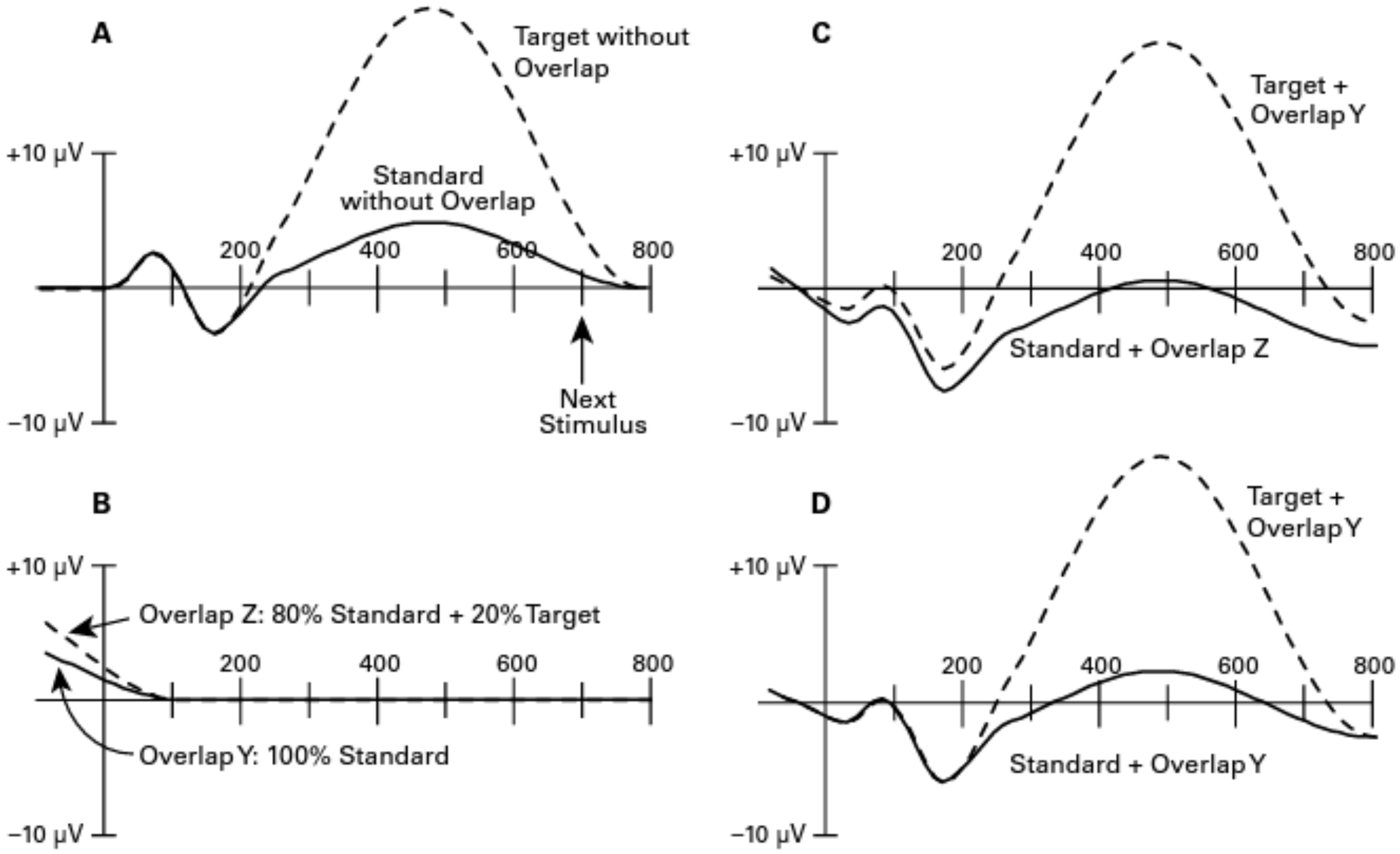


Figure 4.4
Example of differential overlap in the Gedankenexperiment, in which a target stimulus was always preceded by a standard stimulus but a standard stimulus was preceded by a target on 20% of trials and by a standard on 80% of trials. (A) ERPs for target and standard trials without any overlap. (B) Overlap from the last portion of the previous trial. The overlap prior to a target (labeled *Overlap Y*) consists of the last portion of a standard trial. The overlap prior to a standard (labeled *Overlap Z*) consists of the last portion of a standard trial mixed with the last portion of a target trial (80% standard and 20% target). (C) ERP waveforms with the overlap added in. Because the waveforms are baseline corrected, the overlap ends up distorting the waveforms (by pushing them downward). (D) Waveforms that would occur if the ERP waveform for both standards and targets included only trials for which the previous stimulus was a standard. Note that the overlap in this example assumes a very short interval between stimuli and that the subsequent trial would also produce overlapping activity that is not shown here.

was sometimes a target when the current stimulus is a standard), and so the waveform is pushed down farther by the overlap when the current stimulus is a standard than when it is a target. You can see this by comparing the P3 elicited by the frequent stimulus in the presence of overlap (panel C), which has a peak amplitude of approximately 0 μV, with the P3 elicited by the frequent stimulus in the absence of overlap (panel A), which has an amplitude of approximately 5 μV. In contrast, the P3 peak amplitude for the rare stimulus is only slightly smaller in the presence of overlap (panel C) than in the absence of overlap (panel A).

Differential overlap often leads to this sort of pattern, in which an artifactual difference between conditions arises very early (e.g., within 50–200 ms of stimulus onset) and then persists for a very long time (see chapter 6 for a discussion of the importance of looking at the baseline

of differences in arousal. This could also change the preparatory activity before each stimulus, producing different baselines for the bright and dim conditions.

The best way to avoid arousal confounds is usually to vary the conditions unpredictably within each trial block rather than having different trial blocks for the different conditions. In our Gedankenexperiment, for example, it would be trivial to randomly intermix the dim and bright stimuli. However, there are some experiments in which it is necessary to test different conditions in different trial blocks. In these cases, it is sometimes possible to ensure that behavioral accuracy is identical across conditions, which typically equates the arousal level (see, e.g., Leonard, Lopez-Calderon, Kreither, & Luck, 2013).

Confounds Related to Noise and the Number of Trials

In ERP research, the term *noise* refers to random variations in the ERP waveform that are unrelated to the brain activity that you are trying to record (e.g., electrical activity from external devices, skin potentials, etc.). In most cases, noise simply adds random variability to our measurements, reducing the probability that real effects in the data will be statistically significant (i.e., reducing *statistical power*). More noise usually means that we need more trials per subject or more subjects per experiment to achieve statistical significance. In some cases, however, noise can *bias* the data in a particular direction, artificially creating the appearance of an effect. Measurements of peak amplitude are particularly problematic in this regard. All else being equal, the peak amplitude will tend to be larger on average in noisier waveforms than in cleaner waveforms. However, the mean amplitude over a given time range (e.g., 400–600 ms) is not biased in this manner. A fuller discussion of this issue is provided in the online supplement to chapter 9.

In our Gedankenexperiment, this issue arises if we try to compare the peak amplitude for the targets with the peak amplitude for the standards, because fewer trials contribute to the target waveforms than to the standard waveforms, thus making the target waveforms noisier than the standard waveforms. This will tend to bias the peak amplitude to be greater for rare than for frequent stimuli. There are two common ways to solve this problem. First, you can measure mean amplitude rather than peak amplitude (which has a number of additional advantages, as will be discussed in chapter 9). Second, you can create an average of a subset of the standard trials so that the same number of trials contributes to the target and standard waveforms (which should equate the noise levels for these two waveforms). Although this is sometimes the best approach, it decreases the signal-to-noise ratio and therefore decreases your power to find significant effects. I find that some people fail to worry about this problem at all and end up using biased measurements, whereas other people worry too much and end up needlessly sacrificing statistical power when they could have solved the problem simply by measuring mean amplitude instead of peak amplitude.

Tips for Avoiding Confounds

The following list distills the confounds that I just described into six tips for designing ERP experiments:

Tip 1 Whenever possible, avoid physical stimulus confounds by using the same physical stimuli across different psychological conditions (i.e., follow the Hillyard principle). This includes "context" confounds, such as differences in sequential order. Difference waves can sometimes be used to subtract away the sensory response, making it possible to compare conditions with physically different stimuli.

Tip 2 When physical stimulus confounds cannot be avoided, conduct control experiments to assess their plausibility. Don't assume that a small physical stimulus difference cannot explain an ERP effect, especially when the latency of the effect is less than 300 ms.

Tip 3 Although it is often valid to compare averaged ERPs that are based on different numbers of trials, be careful in such situations and avoid using peak-based measures.

Tip 4 Avoid comparing conditions that differ in the presence or timing of motor responses. This can be achieved by requiring responses for all trial types or by comparing subsets of trials with equivalent responses.

Tip 5 To prevent confounds related to differences in arousal and preparatory activity, experimental conditions should be varied within trial blocks rather than between trial blocks. When conditions must be varied between blocks, arousal confounds can be prevented by equating task difficulty across conditions.

Tip 6 Think carefully about stimulus timing so that you don't contaminate your data with offset responses or overlapping activity from the previous trial.

Advice about Timing: Duration, SOA, ISI, and ITI

Almost every experimental design requires you to make decisions about the timing of the stimuli, including durations, SOAs, ISIs, and ITIs. These decisions are often based on the specific goals of an experiment, but there are some general principles that apply to most experiments.

As described earlier, you will usually want to choose a stimulus duration that avoids offset responses, either by choosing a duration that is so short that it doesn't produce a substantial offset response (e.g., 100–200 ms for visual stimuli) or is so long that the offset response occurs after the ERP components of interest (e.g., 1000 ms). In behavioral experiments, it is common for the stimulus to offset when the subject makes a behavioral response to the stimulus. This can sometimes be a good approach in ERP experiments, but it could lead to problems if you decide to look at the postresponse period in response-locked averages (because the sensory offset response will be visible in the response-locked waveforms).

My typical approach for experiments with simple visual stimuli is to use a duration of 100 ms for college student subjects and 200 ms for subjects with poorer perceptual or cognitive abilities. For simple auditory tones, I would typically use a duration of 50–100 ms, including 5-ms rise and fall times. For experiments with more complex auditory or visual stimuli, I typically use a duration of 750–1000 ms.

Determining the optimal amount of time between trials requires balancing several factors. On the one hand, you want to keep the amount of time between trials as short as possible to get the maximal number of trials in your experimental session, thereby maximizing the signal-to-noise ratio of your average ERP waveforms. On the other hand, several factors favor a slower rate of stimulus presentation. First, sensory components tend to get smaller as the SOA and ISI decrease, and this reduction in signal might outweigh the reduction in noise that you would get by averaging more trials together. Second, if the subject is required to make a response on each trial, it becomes tiring to do the task if the interval between successive trials is very short. Third, a short SOA will tend to increase the amount of overlapping ERP activity (which may or may not be a problem, as described earlier). However, using a very long interval between stimuli to minimize overlap may lead to a different problem; namely, anticipatory brain activity prior to stimulus onset (especially if stimulus onset time is relatively predictable).

My typical approach is to use an SOA of 1500 ms for experiments in which each trial consists of a single stimulus presentation and the subject must respond on each trial (e.g., an oddball experiment). I typically reduce this to 1000 ms if the subject responds on only a small proportion of trials, and I might increase it by a few hundred milliseconds for more difficult tasks or for groups of subjects who respond slowly. Of course, there are times when the conceptual goals of the experiment require a different set of timing parameters, but I find that this kind of timing is optimal for most relatively simple ERP experiments.

I almost always add a temporal jitter of at least ±100 ms to the SOA, which avoids the possibility that alpha oscillations will become time-locked to the stimuli and also helps filter out overlapping activity from the previous trial. In a typical oddball experiment, for example, I would use an SOA of 1400–1600 ms. When someone specifies a range like this, it almost always means that the range is broken up into very small increments (e.g., increments of a single screen refresh cycle), and each possible increment in the range is equally likely. For example, with a typical refresh rate of 60 Hz, the time between two stimuli is always a multiple of 16.67 ms, and a range of 1400–1600 ms means that the SOA is equally likely to be 1400.00 ms, 1416.67 ms, 1433.33 ms, . . . 1600 ms. This is called a *rectangular distribution* of SOAs (see chapter 11 for a definition of probability distributions).

Examples from the Literature

Concrete examples can be helpful in clarifying the general principles described in this chapter. To read three of my favorite examples, see the online supplement to chapter 4.

Suggestions for Further Reading

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature, 395*, 597–600.

Gratton, G., Coles, M. G. H., Sirevaag, E. J., Eriksen, C. W., & Donchin, E. (1988). Pre- and post-stimulus activation of response channels: A psychophysiological analysis. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 331–344.

Handy, T. C., Solotani, M., & Mangun, G. R. (2001). Perceptual load and visuocortical processing: Event-related potentials reveal sensory-level selection. *Psychological Science, 12*, 213–218.

Hillyard, S. A., & Münte, T. F. (1984). Selective attention to color and location: An analysis with event-related brain potentials. *Perception and Psychophysics, 36,* 185–198.

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.

Miller, J., & Hackley, S. A. (1992). Electrophysiological evidence for temporal overlap among contingent mental processes. *Journal of Experimental Psychology: General, 121*, 195–209.

Paller, K. A. (1990). Recall and stem-completion priming have different electrophysiological correlates and are modified differentially by directed forgetting. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*, 1021–1032.

Van Petten, C., & Kutas, M. (1987). Ambiguous words in context: An event-related potential analysis of the time course of meaning activation. *Journal of Memory & Language, 26*, 188–208.

van Turennout, M., Hagoort, P., & Brown, C. M. (1998). Brain activity during speaking: From syntax to phonology in 40 milliseconds. *Science, 280*, 572–574.

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.

Winkler, I., Kishnerenko, E., Horvath, J., Ceponiene, R., Fellman, V., Huotilainen, M., Naatanen, R, & Sussman, E. (2003). Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences, 100*, 11812–11815.

Woldorff, M., & Hillyard, S. A. (1991). Modulation of early auditory processing during selective listening to rapidly presented tones. *Electroencephalography and Clinical Neurophysiology, 79*, 170–191.