3
Event-related potentials

The event-related potential technique

We now turn to specific methods for using electrophysiological recordings to investigate cognitive processes. As we alluded to in Chapter 2, there are many approaches for analyzing brain activity and for relating it to other measures of interests (e.g., performance in a memory experiment). The calculation of event-related potentials (ERPs), however, has been such a dominant approach, that sometimes the term "ERP" is used interchangeably with "EEG" (e.g., by referring to an "ERP study" as if the ERP were the dependent measure rather than the EEG activity from which it is derived). The first reports of ERPs (then referred to as "evoked potentials")[1] in conscious humans appeared only 10 years after Hans Berger's publication describing the EEG in humans (Luck, 2014; Nisar & Yeap, 2014; Davis, 1939) and to this day ERPs remain the most popular approach for relating EEG data to performance in laboratory experiments.[2]

Figure 3.1 summarizes the ERP method. EEG activity surrounding a given event is made up of signal reflecting processing of that event as well as background activity ("noise") not related to the event. The goal is to separate signal from noise by sampling EEG activity relative to repeated exposures to the event of interest—the ERP is the average of the resulting time series. As a hypothetical example, imagine a recognition memory experiment of the sort illustrated in Figure 1.1 and suppose that processing of a probe item gives rise to three distinct deflections of the EEG activity (perhaps related to processing the visual features of a probe word, processing the meaning of the word, and initiating the process of determining whether the item had been studied, respectively). Imagine further that three distinct deflection of the EEG activity precede the execution of a response (perhaps these could correspond to the determination that it is time to execute a response, the finalization of the decision which response to make, and the initiation of a motor plan to execute the response). Figure 3.2 illustrates simulated background activity for a single trial i from such an experiment (n_i), simulated signals related to the stimulus and response onset on this trial (s_i), and the corresponding simulated EEG activity recorded at the scalp obtained by adding the signal to the background activity ($x_i = n_i + s_i$). Of course, in practice we will not be able to directly observe the background and signal activity. This example is somewhat contrived, but serves to illustrate the basic issue: on a single trial the underlying signal is so distorted by

[1] The first investigations of ERPs concerned early potentials that were thought to reflect exogenous (evoked) components, i.e., responses to the physical properties of a stimulus. With time, it became clear that internal states could profoundly affect brain activity in response to external events, leading researchers to propose additional endogenous (invoked) components reflecting psychological states (van Boxtel, 1998; Donchin, Ritter, & McCallum, 1979). The term "evoked potential" is still used, especially in clinical settings, to refer to a set of very characteristic potentials in the first ≈ 80 ms following a stimulus (Hillyard & Kutas, 1983). The distinction between evoked and invoked potentials, however, is not always clear, and we thus use the more neutral term ERP throughout.

[2] Our focus here is on EEG, but the same analysis approach is also widely used in magnetoencephalography (MEG) research where the resulting waveforms are known as "event-related fields" (ERFs). Beyond the laboratory, this approach also has a wide range of clinical applications and has even been applied to non-physiological time-series (e.g., event-related analyses of stock prices are known as "event studies"; MacKinlay, 1997).

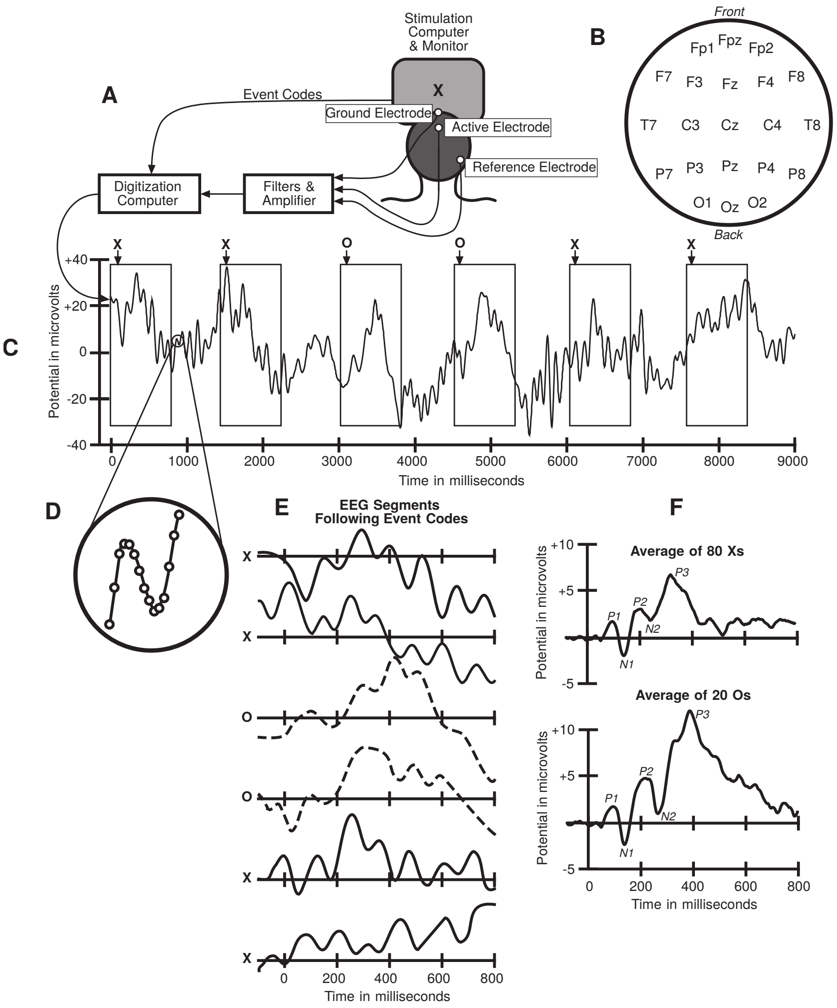

Figure 3.1: Example ERP experiment using the oddball paradigm. The subject viewed frequent Xs and infrequent Os presented on a computer monitor while the EEG was recorded from several active electrodes in conjunction with ground and reference electrodes (A). The electrodes were placed according to the International 10/20 System (B). Only a midline parietal electrode (Pz) is shown in panel A. The signals from the electrodes were filtered, amplified, and then sent to a digitization computer to be converted from a continuous analog signal into a discrete set of digital samples (D). Event codes were also sent from the stimulus presentation computer to the digitization computer, marking the onset time and identity of each stimulus and response. The raw EEG from the Pz electrode is shown over a period of 9 s (C). Each event code during this period is indicated by an arrow along with an X or an O, indicating the stimulus that was presented. Each rectangle shows a 900-ms epoch of EEG, beginning 100 ms prior to the onset of each stimulus. These epochs were extracted and then lined up with respect to stimulus onset (E), which is treated as 0 ms. Separate averages were then computed for the X and O epochs (F). Figure and caption from Luck (2014).

the ongoing background activity that it is effectively invisible in the EEG recording.

We refer to the deflections in the signal (**s**) as "components" and the goal of the ERP method is to estimate their properties from the recorded EEG activity (**x**). Our hypothetical example includes three components that relate to the processing of the probe item. As such, their timing should correlate strongly with the onset of the probe item but weakly with the onset of the subject's response. We further supposed three response-related components whose timing should be relatively invariant with respect to the response, but highly variable with respect to probe onset. To illustrate the power of the ERP method, we simulated a large number of such trials and computed stimulus-locked as well as response-locked ERPs (Figure 3.3). By averaging epochs locked to either stimulus or response onset, we can take advantage of the fact that background noise will cancel and estimate the corresponding components from the resulting ERP waveforms.

For this simulation, we generated random background activity and assumed that the signal associated with the three stimulus-related and the three response-related components consists of deflections made up of half

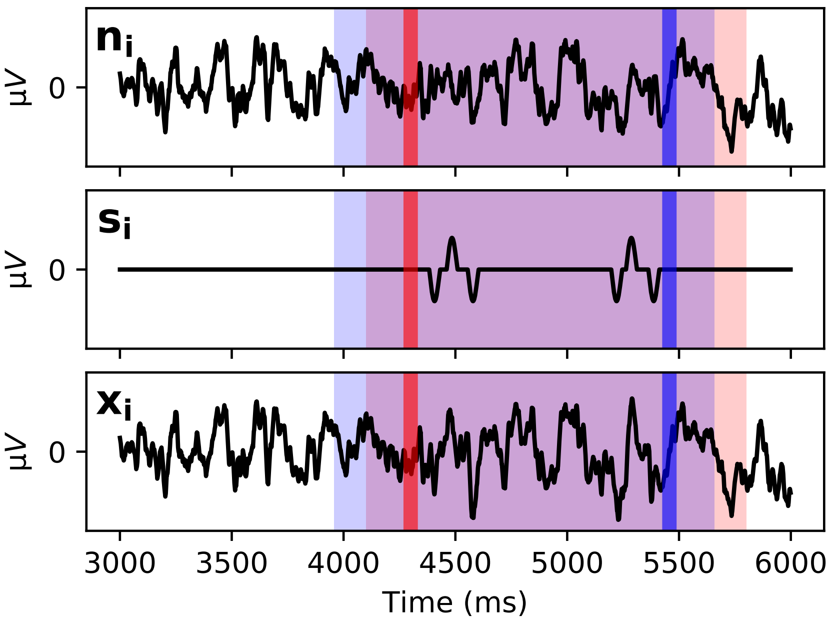

Figure 3.2: Simulated data illustrating the ERP method. Three seconds of simulated background activity and signal combine to form the observed EEG activity. Red and blue vertical lines indicate times of stimulus onset and response execution respectively. Data making up the corresponding event-related vectors $\mathbf{n}_i$, $\mathbf{s}_i$, and $\mathbf{x}_i$ are highlighted. In this example the stimulus-locked vectors contain samples from 200 ms before stimulus onset to 1500 ms after stimulus onset whereas the response-locked vectors contain data from 1500 ms before response execution until 200 ms after response execution.

a cycle of a sine wave with the first and last components causing a negative deflection and the middle component causing a positive deflection (see Figure 3.2). We assumed that response times varied uniformly between 300 and 1500 ms and further assumed that timing of the component most proximal to the respective event (stimulus onset for stimulus-related components and response onset for response-related components) varied uniformly between 60 and 100 ms relative to that event and that the timing of more distal components each independently varied in the same way relative to the timing of the previous component. Clearly none of these assumptions are particularly plausible, except for the fact that we expect some variability in biological systems. Despite the deliberate simplicity of our simulation, it serves to illustrate some basic properties of the ERP method.

Figure 3.3 shows the ERPs obtained by averaging various numbers (as indicated in the middle of each row) of stimulus-locked and response-locked epochs of simulated EEG activity. As the number of trials used to compute each ERP increases, a pattern of two negative deflections separated by a positive deflection emerges. The quality of the signal initially increases quite rapidly, but the increasing number of trials clearly has diminishing effects (note that the number of trials doubles from each row in Figure 3.3 to the next). This is a general property of the averaging procedure (discussed in more detail below) and important to keep in mind when attempting to minimize noise in the generation of ERPs.

Even though each epoch contained both a stimulus and a response, the response-locked components are not apparent in the stimulus-locked ERPs and neither are the stimulus-locked components in the response-locked ERPs. This illustrates that components whose variability is large relative to their extent will be lost in the averaging procedure (because across trials they will get averaged with adjacent components). The relatively small variabil-

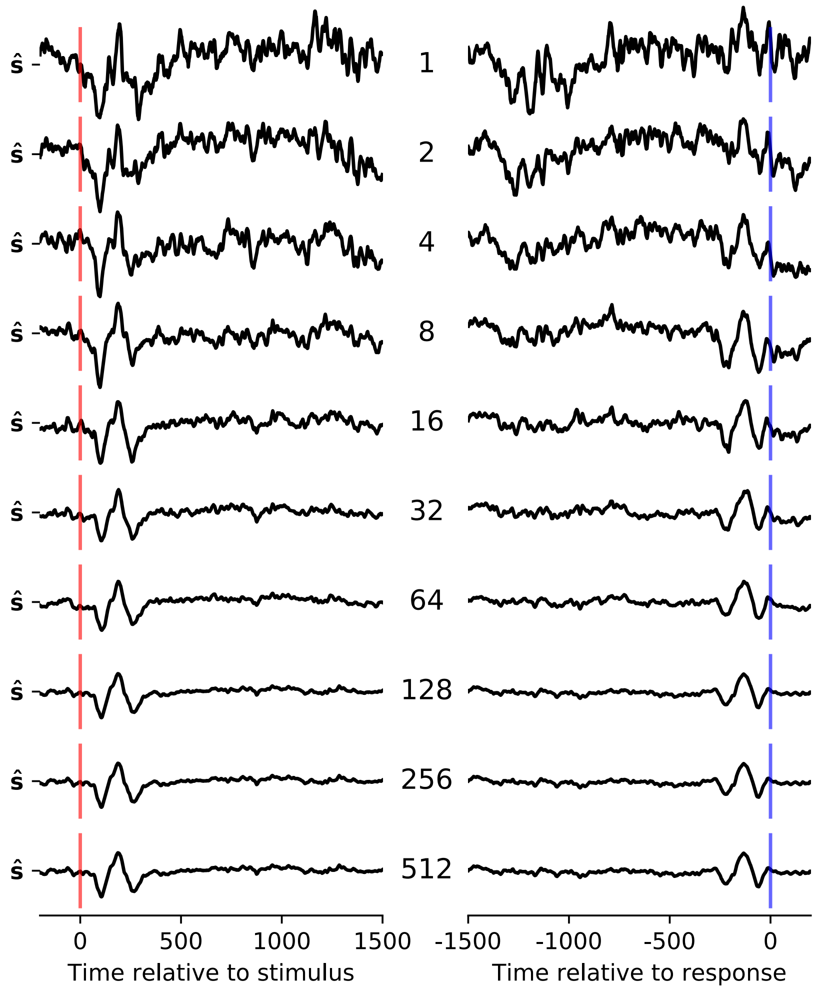

Figure 3.3: Simulated stimulus-locked and response-locked ERPs. Simulated stimulus-locked (left) and response-locked (right) ERPs obtained by averaging various numbers (indicated in the middle of each row) of simulated trials of the sort illustrated in Figure 3.2. Red and blue vertical lines indicate times of stimulus onset and response execution respectively. Note that the number of averaged trials doubles from each row to the next. ŝ indicates the estimated signal (i.e., ERP). See text for details of the simulation.

ity of the components that are related to the locking events results in them causing peaks and troughs in the ERP, but the resulting waveform is a distorted reflection of the underlying components. For example the fact that the timing of each simulated component varied independently with respect to the timing of the next more proximal component (or relative to the event offset in case of the most proximal component) means that the the variability of the components increased with distance from the locking event (because the joint variance of two independently varying events is equal to the sum of the individual variances). This results in a slightly attenuated amplitude and wider spread of the waveforms corresponding to the most distal component. Likewise, individual signals contained gaps between the negative and positive deflections, but these gaps do not appear in the average. Given that biological systems are inherently variable, the ERP waveforms can at best approximate the properties of the underlying components. Real ERPs exhibit fluctuations at relatively low frequencies with peaks and troughs growing increasingly broad as a function of time. Presumably this reflects similar processes with high-frequency fluctuations getting lost in the averaging due to their variability and later components exhibiting relatively larger temporal variability than earlier ones. As we will discuss in more detail below, despite these limitations, ERPs reflect a remarkable number of psychologically relevant variables. In light of the above, it is however important to distinguish between ERP components and the ERP waveforms (which are a noisy reflection of those components surviving the averaging procedure), when using ERPs to inform psychological theory.

A formal background to the ERP method

The averaging procedure of the ERP method will only lead to a completely faithful representation of the underlying components under the following assumptions (Glaser & Ruchkin, 1976):

Linear combination of signal and noise: Just as the sound waves emitted by two speakers in the same room sum together in an audio recording, the assumption is that signal and noise simply sum together (rather than interact). Violations of this assumptions could arise from recording equipment, for example if the amplifier limits the signal to a maximum value which gets assigned to any sample exceeding this threshold.[3]

Invariance of the signal: The signal must be identical for each repetition of the event for the average ERP to reflect the processing of individual events accurately. This assumption could be violated if participants habituate to an event, if the event is processed with variable latency and/or speed, or if the processes elicited by the event vary (e.g., a previously studied item used as recognition memory probe might trigger an elaborate recollective experience or more diffuse feelings of familiarity).

Noise is irregular: With respect the the event of interest, the contributions of noise to the recorded EEG activity must be irregular enough to be indistinguishable from independent samples of a random process. Events that occur at regular intervals could appear at consistent phases of oscillatory background activity, violating this assumption.

As the examples listed with each assumption (as well as our discussion above) suggest, it is unrealistic to hope that these assumptions are not violated in practice. Furthermore without independent means to distinguish between signal and noise, it is difficult to test these assumptions. However, careful experimental design can mitigate some potential problems (e.g., EEG studies usually contain randomly jittered delay periods to prevent the entrainment of background activity to experimental events) and, as we have discussed above, to the extent that violations are minor, ERPs will still provide useful information (Glaser & Ruchkin, 1976).

The EEG activity at a given channel (e.g., a pair of electrodes) can be expressed as a vector, $\mathbf{x}$, holding the recorded voltage samples in the order in which they were recorded. The following formal introduction to the ERP method and its properties follows closely that of Glaser and Ruchkin (1976) who provide additional details and derivations.

To generate an ERP, we partition the EEG activity to obtain a vector of EEG activity for each repetition, i, of the event of interest and assume that each such vector, $\mathbf{x}_i$, represents the sum of the signal, $\mathbf{s}$, and noise, $\mathbf{n}_i$: $\mathbf{x}_i = \mathbf{s} + \mathbf{n}_i$ (as indicated above, the signal is assumed to be invariant across repetitions of the event and thus does not have subscript indicating a par-

[3] This is known as "signal clipping".

ticular instance of the event). The goal is to compute an estimate of $\mathbf{s}$, $\hat{\mathbf{s}}$, by averaging across the N repetitions of the event, yielding

$$\hat{\mathbf{s}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \mathbf{s} + \frac{1}{N}\sum_{i=1}^{N}\mathbf{n}_i \; . \tag{3.1}$$

Equation 3.1 defines the averaging at the heart of the ERP method. Without loss of generality we can assume that the expected value of the noise, $E\left[\mathbf{n}_i\right]$, is $\mathbf{0}$. We demonstrate that $\hat{\mathbf{s}}$ is an unbiased estimator of $\mathbf{s}$ by showing that its expected value, $E\left[\hat{\mathbf{s}}\right]$, is equal to $\mathbf{s}$:

$$E\left[\hat{\mathbf{s}}\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\right] = \mathbf{s} + \frac{1}{N}\sum_{i=1}^{N}E\left[\mathbf{n}_i\right] = \mathbf{s}$$

The above discussion makes it clear that the precision of $\hat{\mathbf{s}}$ depends on the relative magnitudes of the signal and the average of the noise. Given that $\hat{\mathbf{s}}$ is an unbiased estimator of $\mathbf{s}$, we can expect that as N increases, $\hat{\mathbf{s}}$ approaches $\mathbf{s}$. We will now explore this relationship between the number of observations and the precision of $\hat{\mathbf{s}}$. The vector of standard errors of the estimated signal (i.e., the ERP), $\sigma_{\hat{\mathbf{s}}}$, contains the noise residual at each sample. Denoting the (diagonal) variance-covariance matrix of the noise, $E\left[\mathbf{n}_i\mathbf{n}_i^T\right]$, by Σ_n (the superscript T denotes the transpose; the off-diagonal elements of this matrix are zero due to the assumed independence of noise samples), we can first express the (also diagonal) variance-covariance matrix of the ERP, $\Sigma_{\hat{\mathbf{s}}}$, as a function of Σ_n and N and then obtain $\sigma_{\hat{\mathbf{s}}}$ by taking the element-wise square root of the main diagonal:

$$\begin{aligned}
\Sigma_{\hat{\mathbf{s}}} &= E\left[(\hat{\mathbf{s}}-\mathbf{s})(\hat{\mathbf{s}}-\mathbf{s})^T\right] \\
&= E\left[\sum_{i=1}^{N}\frac{\mathbf{n}_i}{N}\frac{\mathbf{n}_i^T}{N}\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}E\left[\mathbf{n}_i\mathbf{n}_i^T\right] \\
&= \frac{1}{N^2}\times N\times\Sigma_n
\end{aligned}$$

Thus

$$\sigma_{\hat{\mathbf{s}}} = \frac{\sigma_n}{\sqrt{N}} \; , \tag{3.2}$$

where $\sigma_n = \operatorname{diag}\left(\Sigma_n\right)^{\circ\frac{1}{2}}$ and $\circ$ indicates that the square root is taken for each vector element.

The values in $\sigma_{\hat{\mathbf{s}}}$ denote the standard error at each sample and will vary to the extent that noise is nonstationary. It is worth taking a moment to reflect on what the above result means for efforts to increase the signal to noise ratio in ERP analyses. Clearly, all else being equal, more data leads to more precise estimates of the signal. The square root in the denominator of Equation 3.2, however, implies that to reduce noise by half, quadruple the amount of data is needed. This relationship highlights the importance of minimizing noise in the recording rather than relying solely on the averaging process to increase the signal to noise ratio (Luck, 2014).

Practical issues

The above example was deliberately simplistic to illustrate the basic issues associated with generating and interpreting ERPs. In practice, components can vary across many dimensions and be sensitive to a wide range of variables including stimulus properties, experience, or task demands. Even with the highly regular background activity and component properties in the simulation, however, we observed distortions in the average waveforms. In real EEG recordings, potentials can drift over time due to factors not related to brain activity which could cause serious distortions in the resulting ERPs if not appropriately countered (Luck, 2014). A generally effective strategy for eliminating this source of noise is the application of a *baseline correction*. The idea is that the period just prior (or, usually in the case of response-locked ERPs, just after) the event should be relatively free of event-related activity. One can thus center individual waveforms such that the average potential of the baseline activity is 0 μV. With this common preprocessing step, the amplitudes of the various deflections in the ERP waveforms represent the average difference from baseline rather than the average absolute potentials.

The inclusion of a (typically 100–200 ms) baseline period is also useful for assessing the success of the averaging procedure in eliminating background activity. To the extent that brain activity during the baseline period is not event-related, variability in the this period reflects noise and any deflections outside the baseline period that do not exceed those in the baseline period are unlikely to represent event-related activity (Luck, 2014; Woodman, 2010).

ERPs are frequently visualized by graphing individual waveforms with time on the abscissa and voltage on the ordinate (see, e.g., Figure 3.1F). It so happened that the first published ERPs (Davis, 1939) were plotted with what would conventionally be viewed as an inverted y-axis (i.e., positive voltages were plotted below zero and increased in the downward direction with voltages decreasing in the upwards direction). This quirk caught on with a large number of ERP researchers, but increasingly the trend is to follow the standard convention of having values increase in the upwards direction (Luck, 2014). As a result, it is important to pay careful attention to to orientation of the y-axis in published ERP waveforms and to clearly indicate the axis orientation when illustrating ERPs.

ERP components

ERP researchers have identified a wide range of components on the basis of systematic variations of ERP waveforms in response to experimental manipulations. Unfortunately naming conventions for these components are somewhat inconsistent. For many components the general pattern is that the name starts with the letter "P" or "N" to indicate a positive or negative deflection in the associated ERP waveform respectively[4] followed by a number that either indicates an ordinal position (as in "N2" for the second negative deflection; see also Figure 3.1F) or an approximate time in ms corresponding to the peak of the deflection (as in "P300" for a positive deflection peaking around 300 ms after the event onset).[5] Other components have more descriptive names such as the "lateralized readiness potential" (LRP), a response-related component that distinguishes which hand will execute a response or the contingent negative variation (CNV) a negative deflection thought to

[4] For magnetic fields recorded in magnetoencephalography experiments, the letter "M" is used instead.

[5] This labeling is not as intuitive as it might appear. The polarity of a component can vary with the placement of electrodes and choice of reference. Additionally, not all components are visible from all electrode sites and thus the ordinal indicator might not correspond to the deflections observed in a specific waveform. Similarly, the timing indicator need not be be very exact, especially for later components (e.g., positive deflections with a peak as late as 600 ms are often labeled as P300). A further complication arises from the fact that some (early) components are modality-specific and thus the label "P1" refers to different components depending on whether it was recorded relative to a visual or an auditory stimulus (Luck, 2014).

index anticipatory processes (Luck, 2014).

Earlier components tend to be highly sensitive to perceptual features of the eliciting stimuli whereas later components tend to covary more strongly with internal states (e.g., those that reflect task demands). One of the most widely studied components of the latter kind is the P300 (also known as P3 or P3b) which we briefly introduced in Chapter 2. One reason for the appeal of this component is that it is sensitive to experimental contingencies that require the categorization of the eliciting stimuli along task-defined dimensions. This makes this component useful for the study of a wide range of cognitive processes and allows it to establish an upper bound on the duration of the processes responsible for this categorization (Luck, 2014). An example is shown in Figure 3.1F where this component (labeled P3 in the figure) is considerably larger for rare stimuli.

In practice the distinction between different components is often tricky, because there can be considerable overlap (and variability) in their timing. Additionally, earlier deflections can have lasting effects on the ERP, making it difficult to unambiguously attribute ERP differences to later components when earlier differences are also apparent. For example, one might want to compare memory for faces with that for names by presenting images of faces and strings of letters spelling out the names on a computer screen with the instruction that these be memorized for a subsequent memory test. Comparing ERPs locked to the presentation of subsequently remembered faces with those locked to the onset of subsequently remembered names will reveal differences associated with the different perceptual properties of these stimulus types (e.g., it is known that that faces elicit an enhanced early negativity known as the N170 component) in addition to any differences associated with memory processes that are sensitive to the stimulus class. It is therefore important to structure comparisons between ERP waveforms such that the process of interest is isolated.

ERPs and human memory

In recognition memory tasks (see Figure 1.1) one can observe ERPs relative to the probe items (just as we proposed in our hypothetical example above). ERP waveforms distinguish between correctly identified old ("hits") and new ("correct rejection") probe items between around 300–500 ms at mid-frontal electrodes (this effect is sometimes called FN400 where the "F" prefix specifies the frontal scalp distribution) and between around 400–800 ms at left-parietal electrodes (see Figure 3.4; Wilding & Ranganath, 2012; Luck, 2014). Dissociations between these two components have prompted interpretations of these two components within the framework of dual-process theories of recognition memory that postulate that two distinct types of evidence drive recognition decisions: *familiarity* and *recollection* (Yonelinas, 2002; Yonelinas, Aly, Wang, & Koen, 2010; Malmberg, 2008). Familiarity refers to the diffuse sense that an item has been studied and the size of the mid-frontal old-new effect (FN400) is often interpreted as an electrophysiological index of this construct. Recollection, on the other hand, indicates the ability to retrieve contextual details associated with the study episode. The size of the left-parietal old-new effect is commonly thought to covary with recollective experience (Curran, 1999; Rugg & Curran, 2007; Wilding & Ranganath, 2012; Luck, 2014). Recent work by some of us, however, suggests that EEG

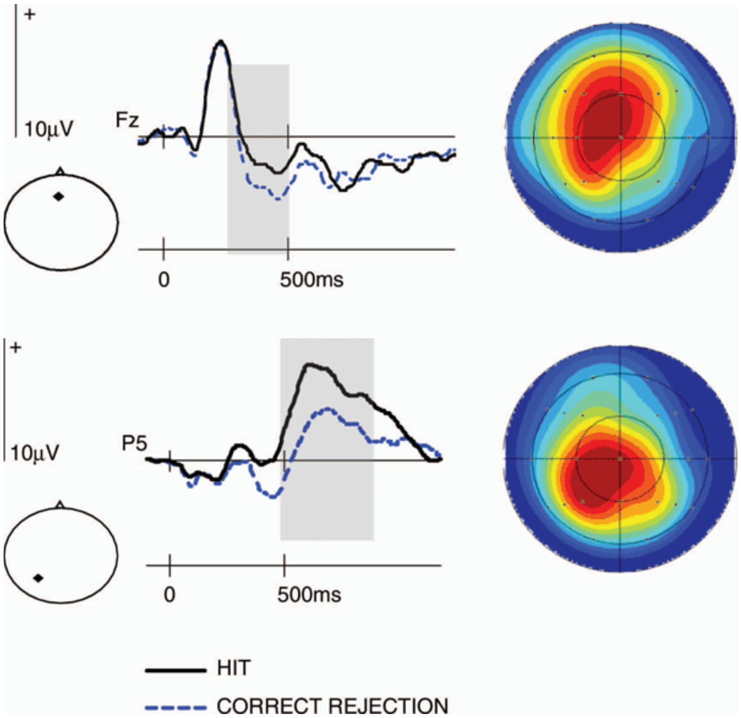

Figure 3.4: ERP old-new effects. An early mid-frontal (sometimes called FN400; top panel) and a late left-parietal (bottom panel) ERP old-new effect exhibiting greater positivity for hits (i.e., old items that were correctly recognized as old; black solid line) than for correct rejections (i.e., new items that were correctly classified as such; blue dashed line). The locations of the electrodes from which the ERPs were obtained (Fz and P5 respectively) are indicated in a small top down schematics of the head (the triangle at the top of the circle is a stylized nose and the dot inside the circle indicates the electrode location). On the right of each panel is a overhead scalp map (top is frontal) indicating the size of the difference between the ERP waveforms for hits and correct rejections across the entire scalp in the time windows highlighted in gray (300–500 ms and 500–800 ms respectively). Red indicates the largest difference and cooler colors indicate smaller differences. These plots confirm that the scalp distribution of the early effect is strongest at mid-frontal electrodes whereas the scalp distribution of the latter effect is more pronounced at left-parietal electrodes. Adapted from Wilding and Ranganath (2012).

activity reflects a unitary recognition signal combining the available sources of evidence distinguishing between old and new probe items (Weidemann & Kahana, 2019).

One can also study the processing of items during study as a function of subsequent memory. Such differences are known as *Dm* ("difference due to memory") or *subsequent memory effects*. Subsequent memory effects in ERPs typically do not manifest as distinct peaks, but instead as sustained positive deflections at centro-parietal electrodes for subsequently remembered items. These differences start around 400 ms after the onset of the study item and last several hundred milliseconds. Additional subsequent memory effects have been observed at left anterior electrodes and can be sensitive to the types of stimuli that are being remembered (Luck, 2014).

Experiments designed to investigate the properties of visual working memory, are particularly well suited to the ERP method. Trials in these experiments typically consist of a brief visual display followed by a fixed retention interval that ends with a test array (see Figure 3.5A).[6] In an influential study Vogel and Machizawa (2004) identified what has become known as the Contralateral Delay Activity (CDA)—a sustained negative deflection of the ERP contralateral to the visual hemifield containing the memory set. In this experiment participants were asked to memorize colored squares in one hemifield of the display and, after a retention interval, were presented with a test array of squares that was either identical to the memory array or included a change to the color of one square in the cued hemifield (see Figure 3.5A). Figure 3.5B shows ERPs for electrodes ipsilateral and contralateral to the cued hemifield and Panels C of this figure indicates that the size of

[6] Chapter **??** presents a detailed discussion of the electrophysiology of working memory.

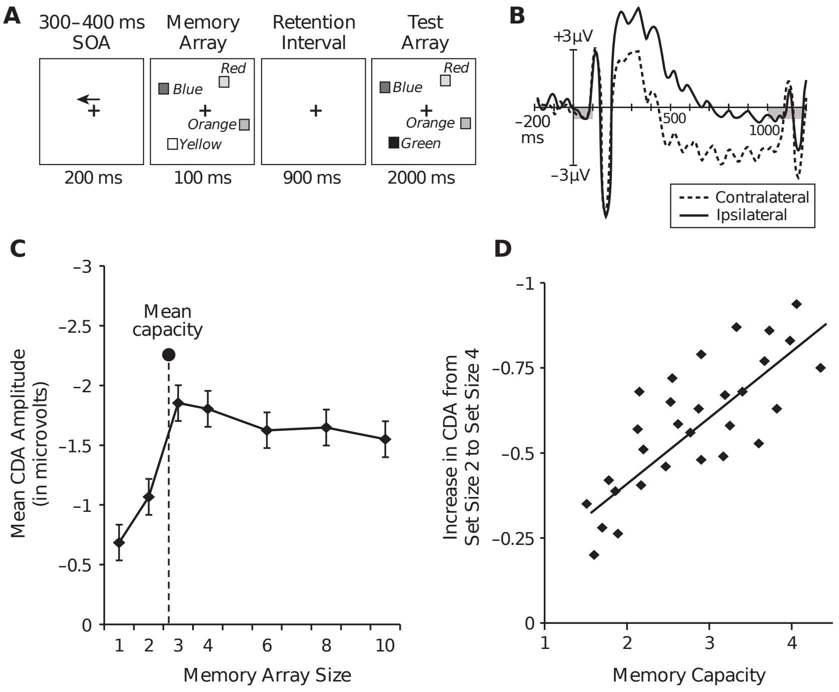

Figure 3.5: Working memory effects in ERPs (Vogel & Machizawa, 2004). (A) Design of the visual working memory task. An arrow cued the side on which square colors were to be memorized. (B) Grand average ERP waveforms for a memory array of size 4 (i.e., 4 items in the cued hemifield were to be remembered, 4 additional items in the uncued hemifield never changed). ERPs were time-locked to the onset of the memory array and averaged across electrodes at lateral occipital and posterior parietal locations. Gray rectangles indicate the presence of the memory and test array respectively. (C) Mean amplitude of the contralateral delay activity (CDA) between 300–900 ms after onset of the memory array as a function of the size of the memory array. Mean visual working memory capacity (2.8 items, estimated from response patterns) is indicated with a dashed vertical line. Error bars indicate 95% confidence intervals. (D) Visual working memory capacity plotted against increase in CDA amplitude between memory arrays of size 2 and 4 ($r = 0.78$). All ERPs and derived measures shown here are based on all trials, but follow-up work suggested that results are very similar if ERPs are only generated for trials with correct responses (Vogel, personal communication, Oct. 4, 2017). Adapted by Luck (2014) from Perez and Vogel (2012). Copyright 2012 Oxford University Press.

the difference between these waveforms rose with the size of the memory set, but only until memory capacity (derived from the patterns of errors in this task) was reached, ruling out that this effect was simply a response to the increased number of displayed squares. Vogel and Machizawa (2004) also related the size of the increase of the CDA between memory sets of 2 and 4 for each individual to the corresponding working memory capacity (derived from each individual's pattern of errors) and found a striking correlation between these measures (shown in Figure 3.5D).

Quantifying properties of ERP waveforms

To allow ERPs to inform psychological theory, it is necessary to quantify how they vary with time and experimental condition. The early parts of ERPs are usually characterized by distinct peaks and troughs and some studies attempt to quantify corresponding amplitudes and/or latency. This is trickier than it might appear. For example, simply selecting the maximum in a time-window could produce spurious results if noise in the waveform causes high amplitude blips (which may not coincide with the peak of the component). Likewise, if the time window is not carefully chosen, the maximum might correspond to a value in the rise towards a subsequent peak rather than representing the summit of the component of interest. More sophisticated procedures for estimating the amplitude and latency of peaks and troughs exists, but it is worth bearing in mind that there is nothing inherently special about these points. Alternative measures, such as when the waveform first deviates from baseline, can be better suited for relating components to underlying processes and, indeed, are often used, especially when the aim is to put an upper bound on the timing of a specific process associated with a given ERP component (for a comprehensive review of different measures to quantify properties of ERP waveforms, see Chapter 9 of Luck, 2014).

A particularly popular measure is the area between the waveform and 0 μV (or between two waveforms) within a given time interval. An important reason for the popularity of the area measure is that it is relatively robust to small fluctuations in the waveforms, but for later components that no longer exhibit clearly defined peaks and troughs, there are few alternatives (see Figure 3.5 for an example of the use of the area measure to study the electrophysiology of working memory).

The role of ERPs in memory research

ERPs are particularly well suited for studying neural activity immediately following (or preceding) well-defined events. Encoding and retrieval processes, however, can extend over relatively long time periods and are not always easy to link to specific events. For example, the presentation of a study item may prompt the retrieval of a previously studied item and deliberate rehearsal processes are difficult to control (see Chapter 1 for a discussion of these issues). As indicated above, ERPs tend to reflect low frequency deflections of the EEG activity, yet spectral features of the EEG activity that are usually lost in the ERP have also been shown to contain information about episodic memory processes (Nyhus & Curran, 2010; Jacobs, Hwang, Curran, & Kahana, 2006).

Despite these limitations, ERPs have been reliably linked to performance in memory tasks and are almost certainly the most popular method in investigations of the electrophysiology of human memory. Undoubtedly part of this popularity is due to the the simplicity of the computations that are required for the generation of ERPs and to the substantial prior literature establishing the properties of ERPs in a wide range of experimental contexts. The rapid increase in the availability of more powerful computational resources, however, has led to an increasing use of methods that investigate neural activity on a trial-by-trial basis and that also consider spectral features that are lost in the generation of ERPs. It is likely that the use of these methods (which will be the foci of subsequent chapters) will continue to gain in popularity and perhaps even surpass the use of ERPs in the coming decades. As, we hope, will become obvious in the course of reading this book, the various methods for investigating neural activity have complementary strengths and weaknesses. A comprehensive understanding of the electrophysiology of human memory is therefore likely to depend on the results of investigations using a wide range of methods and the aim of this and the coming chapters is to highlight those that have to date been proven to be most promising for this endeavor. Before we turn to other univariate methods for analyzing brain activity, however, we provide a brief overview over statistical issues that arise in the analysis of electrophysiological data and the main methods for addressing them.

Statistical analyses of electrophysiological data

Several practical issues arise when determining the parameters for the analysis of ERP waveforms and other types of electrophysiological data. For example, to determine the time window(s) for which the area under an ERP waveform should be computed, it might be tempting to plot the ERPs, and to determine the time window(s) on the basis of the shape of these waveforms (e.g., if the difference between the ERPs for two experimental conditions appears particularly large between 450 and 650 ms after event onset, one might wish to calculate statistics on the area between the corresponding ERPs within these two time points). Until not very long ago, this approach was quasi-accepted practice and, unfortunately, the associated problems are still not universally appreciated. In short, the issue with picking analysis parameters on the basis of explorations of the data that are to be analyzed is that it severely inflates the probability of finding an effect when none is present (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Kilner, 2013).[7] In null hypothesis significance testing (NHST), the p-value indicates the probability of obtaining the observed results (or more extreme results) given that the null hypothesis is true.[8] The validity of this p-value, however, depends on unbiased sampling of the data. In cases where the time window (or other analysis parameters) are informed by explorations of the same data set, the probability of observing an effect that is at least as large as the observed one, given that the null hypothesis is true, may be substantially larger than the p-value indicates.

One way to avoid this problem is to determine analysis parameters *a priori*, for example on the basis of an independent data set. A complementary approach is to test multiple analysis parameters, for example by testing multiple time windows (determined *a priori*!) or doing away with time windows altogether and calculating separate statistics for each sample. Because of the complexity of electrophysiological data sets (for each event, many samples are often collected across a large range of sensors), such data sets are routinely subjected to multiple statistical tests. This raises related statistical issues which we will discuss next.

Multiple comparisons

In NHST, a threshold, α, is set (most commonly to $\alpha = 0.05$) such that results with p-values that fall below this threshold are deemed statistically significant (i.e., the null hypothesis is rejected). In this framework one can make two types of errors (see Table 3.1): A Type I error refers to erroneously rejecting the null hypothesis when there is no effect. In contrast, a Type II error refers to an erroneous failure to reject the null hypothesis when an effect is present.[9] The α-level directly controls the Type I error rate for a given statistical test, but, all else being equal, reducing the Type I error rate increases the Type II error rate and vice versa. When conducting multiple statistical tests, it is usually desirable to either control the probability that any of the tests leads to a Type I error (this is the *family-wise error rate*, FWER) or to limit the total proportion of Type I errors across all tests (this is the *false discovery rate*, FDR).

	Fail to reject H_0	Reject H_0
H_o is true	✓	Type I error (α)
H_0 is false	Type II error (β)	✓

Table 3.1: Decision table for an individual null hypothesis significance test. The significance threshold α controls the Type I error rate. The inverse of the Type II error rate $(1 - \beta)$ is known as the power of a statistical test. For multiple statistical tests controlling the FWER limits the probability of making at most 1 Type I error across the family of statistical tests whereas controlling the FDR limits the expected proportion of Type I errors among all significant results (i.e., the proportion of erroneous rejections of H_0 among all rejections of H_0 across the family of statistical tests).

To illustrate the issue, imagine attempting to rappel from a burning building using a makeshift rope made out of bed sheets. If you knew that each knot joining two bed sheets independently had a 5% chance of failure, would you prefer using 2 long bed sheets or 4 short bed sheets to make the rope for your escape? With no other potential points of failure, it should be clear that a rope with fewer knots is the safer escape option. In this case, assuming a failure of the rope is fatal, you will have a 95% chance of making it out alive with 1 knot, but a less than 86% ($0.95^3 \times 100$) chance of survival with 3. With

[7] This approach to data analysis is reminiscent of the proverbial Texas sharpshooter who shoots at a wall and then paints a bull's-eye around the bullet hole. A related concept is that of *overfitting* which we will discuss in detail in Chapter **??**.

[8] If this definition of the p-value surprises you, you may have been led astray by ubiquitous erroneous definitions that sometimes even make it into text books and other academic writing. The p-value is emphatically not the probability that the null hypothesis is true, an index of the likelihood that the results will (fail) to replicate, or some variation on these or similar themes.

[9] This situation is analogous to that in recognition memory tests where one can erroneously identify a new item as old or fail to recognize an old item as old. Rather than referring to these errors as Type-I and Type-II errors respectively, they are usually labeled as "false alarms" and "misses" in the context of recognition memory tests or other binary choice tasks.

14 knots or more, the rope is more likely to fail than to hold despite each individual knot only having a 5% failure rate. Likewise, imagine calculating ERPs for 2 experimental conditions and running separate statistical tests for the difference between them at each sample across participants. Data from a 1 s period sampled at 1000 Hz would produce a FWER very close to 1.0 ($1 - 0.95^{1000}$ to be exact) if the tests were independent. Below we will discuss different methods for controlling the FWER or the FDR across multiple statistical tests. With the exception of the Bonferroni correction, which is not generally used when correcting for large numbers of tests, these methods capitalize on dependencies between the statistical tests[10] to limit either FWER or FDR without disproportionately increasing the Type II error rate.

Many experimental designs that include multiple independent variables are routinely analyzed within an Analysis of Variance (ANOVA) framework. A common misconception is the idea that the use of an ANOVA instead of, say, calculating individual *t*-tests avoids such issues with multiple comparisons. It is true that the ANOVA controls Type I error levels for each statistical test so that the number of factor levels does not affect the Type I error level for the test of the corresponding main effect. However, in ANOVAs we typically compute multiple statistical tests and the Type I error rate for this family of tests is not controlled. An ANOVA with 2 factors usually involves tests of the 2 main effects and a test of their interaction for a total of 3 statistical tests. The addition of just 1 additional factor more than doubles the number of statistical tests to 7 (3 main effects + 3 2-way interactions + 1 3-way interaction). When ANOVAs are used in analyses of electrophysiological data, it is not uncommon to add additional factors (e.g., for electrode locations across the anterior-posterior and/or the left-right dimensions, or for multiple time windows).[11] With 4 factors (corresponding to a total of 15 statistical tests) or more, the FWER exceeds 0.5 making it more likely than not that at least one of these tests produces a significant result even when all null hypotheses are true. It is therefore important not to be fooled into believing that issues of multiple comparisons do not apply when computing "only one" ANOVA. Especially for complex ANOVA designs with 4 or more factors, it can be prudent to apply the kinds of controls of the FWER or FDR that we introduce below (see Bishop, 2014, for a more thorough discussion of these issues).

Bonferroni and Holm correction to control the FWER

The simplest way to control the FWER is to divide the uncorrected significance threshold α by the number of statistical tests, m, to compute a corrected threshold value α/m on which the statistical decisions are based; this procedure is known as the Bonferroni correction. For 100 statistical tests and $\alpha = .05$, the corrected significance threshold would be $0.05/100 = 0.0005$ and only tests with p-values below this threshold would be considered significant at the (family-wise) .05 threshold.[12] The Bonferroni correction does not consider dependencies between the statistical tests and is therefore very conservative in situations such as those discussed in this chapter. This reduces the power to detect effects in the data, especially among a large number of tests, and therefore the Bonferroni correction is not typically used for electrophysiological data.

A less conservative alternative to the Bonferroni correction is a procedure

[10] In electrophysiological data, measures are usually highly correlated across nearby time points and sensors introducing substantial dependencies between associated statistical tests.

[11] Such use cases usually violate the ANOVA assumption that samples at different factor levels are independent from each other.

[12] Proof that the Bonferroni correction controls the FWER: Suppose that we are conducting a total of m statistical tests. Let p_i, $i \in \{1, \ldots, m\}$ be the corresponding p-values and I_0 the set of indices of the true null hypotheses. The FWER is the probability of making at least one Type I error,

$$\begin{aligned} \text{FWER} &= P(\bigcup_{I_0}\{p_i \leq \alpha/m\}) \\ &\leq \sum_{i \in I_0} P(p_i \leq \alpha/m) \\ &\leq |I_0| \times \alpha/m \leq \alpha \end{aligned}$$

developed by Holm (1979). For this correction, all p-values are ordered from smallest to largest and one finds the smallest index, i in the resulting list of p-values for which $p_i > \alpha/(m-i+1)$, where m is the number of statistical tests. The null hypotheses corresponding to the p-values at index $i-1$ or below are rejected. This procedure can also be used to calculate adjusted p-values ($\tilde{p}$) as follows (Yekutieli & Benjamini, 1999):

$$\tilde{p}_i = \max_{k=1,\ldots,i} \Big\{ \min((m-k+1) \times p_k, 1) \Big\}.$$

Table 3.2 summarizes the threshold values at each step of both FWER control procedures. Even though Holm's (1979) procedure is less conservative than the Bonferroni correction, it still leads to very low significance thresholds when the number of statistical tests is large (see Figure 3.6 for an illustration) and thus limits the power to find effects. In practice it is therefore often better to limit the FDR instead of the FWER or to use non-parametric shuffling procedures to limit the FWER, both of which we will discuss below.

Bonferroni	Holm
α/m	α/m
α/m	$\alpha/(m-1)$
$\vdots$	$\vdots$
α/m	α

Table 3.2: Table of significance thresholds at each step of two FWER control procedures. p-values are sorted from lowest to highest and are compared to the corresponding significance thresholds at each step.

Controlling the False Discovery Rate

In situation where a single Type I error is less problematic than the failure of a single knot in the hypothetical scenario sketched out above, controlling the FDR instead of the FWER can often represent a good compromise between limiting the number of both Type I and Type II errors. Controlling the FDR constitutes *weak control* of the Type I error rate, because, on average, we would expect 5% of the statistically significant tests to result in a Type I error. Thus FWER control implies FDR control (but not vice versa) and when all null hypotheses are true, controlling the FWER and the FDR are equivalent (Dudoit, Shaffer, & Boldrick, 2003).

The two main methods for controlling the FDR are that by Benjamini and Hochberg (1995) and that by Benjamini and Yekutieli (2001). The former is less conservative (and more popular), but has been shown to control the FDR only when the individual tests are not negatively correlated (Benjamini & Yekutieli, 2001). For this procedure p-values are sorted from smallest to largest and the largest index i in the corresponding list of sorted p-values for which $p_i \leq (i/m) \times \alpha$ is identified. The null hypotheses at this and lower indices are rejected. Adjusted p-values ($\tilde{p}$) are calculated as follows (Yekutieli & Benjamini, 1999):

$$\tilde{p}_i = \min_{k=i,\cdots,m} \Big\{ \min(m/k \times p_k, 1) \Big\}$$

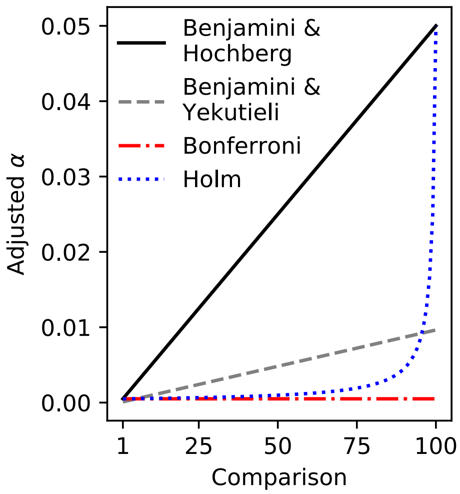

Figure 3.6: Threshold functions under different multiple comparison procedures. Significance threshold are shown for each of 100 tested hypothesis (sorted from lowest to highest p-value). The threshold for the Bonferroni correction is constant across comparisons, whereas thresholds for the FDR control procedure proposed by Benjamini and Hochberg (1995) increase linearly with threshold. Corresponding thresholds for the other procedures fall between these two functions.

The assumption that statistical tests are not negatively correlated can be problematic. For example, some electrodes might be placed at opposite ends of a dipole, which could lead positive deflections at one electrode to coincide with negative deflections at the other. An alternative method to control the FDR that is valid even for these cases is to identify the largest index i in the list of sorted p-values for which $p_i \leq (i/(m \times \sum_{j=1}^{m} 1/j) \times \alpha$ (Benjamini & Yekutieli, 2001). As with the other procedure, null hypotheses at this and lower indices are rejected. The adjusted p-values ($\tilde{p}$) for Benjamini and Yekutieli's (2001) FDR correction is as follows:

$$\tilde{p}_i = \min_{k=i,\cdots,m} \Big\{ \min(\sum_{j=1}^{m} \frac{1}{j} m/kp_k, 1) \Big\}$$

Table **??** summarizes the significance thresholds for these two FDR correction procedures. These thresholds differ by the harmonic sum term $\sum_{j=1}^{m} 1/j$, which is approximately $\log(m)$ when m is large. In practice, the method by Benjamini and Yekutieli (2001) is seldom used, because it leads to substantially higher levels of Type II errors than that proposed by Benjamini and Hochberg (1995). Furthermore, Benjamini and Hochberg's procedure appears to be relatively robust to violations of the assumptions that tests are not negatively correlated (Clarke & Hall, 2009; Groppe, Urbach, & Kutas, 2011). Figure 3.6 shows significance thresholds for each comparison across the different methods for controlling the FWER and the FDR discussed so far.

Permutation Tests

Above we provided the definition of a p-value as the probability of obtaining the observed (or more extreme) results given that the null hypothesis is true. The idea of permutation tests is to calculate this probability non-parametrically (i.e., without making any assumptions about the shape of the sampling distribution) by generating many new samples from the observed data under the assumption that the null hypothesis holds. Imagine observing the area under a specific portion of the ERP waveform in 10 individuals who were each subjected to two experimental conditions. If the null hypothesis that there is no difference between the experimental conditions is true, we can permute the condition labels to generate many hypothetical data sets that come from the same distribution as the actually observed data. For each individual, there are two possible assignments of condition labels to conditions (the correct assignment, and the permuted assignment) leading to a total of 1024 (2^{10}) possible permutations, one of which corresponds to the actually observed data set (this permutation test is a non-parametric analog to a paired t-test, taking into account that every individual provides data for both conditions). For each permutation we could calculate a t-statistic of the difference in area under the two ERP waveforms corresponding to the respective condition labels. If the null hypothesis is true, the difference in areas for the actually observed condition labels should come from the same distribution as the differences for the permuted condition labels and we would expect the t-statistic for the actually observed difference to fall near the center of the distribution of t-statistics corresponding to the differences for the permuted condition labels. The p-value for this permutation test simply corresponds to the proportion of t-statistics in this distribution that are at least as extreme as the actually observed t-statistic. Thus, if the actually observed t-statistic is among the $\alpha \times 100\%$ most extreme t-values, we can reject the null hypothesis that the areas for the two experimental conditions are identical.

With larger number of observations, it can be impractical to compute the distribution of the relevant statistic for each possible permutation. It is reasonable in these cases to instead use a large number of random permutations. Whatever the number of permutation, the corresponding p-vale is determined by the rank of the actually observed statistic within the permutation distribution and thus the number of permutations determines the resolution of the resulting p-value: For n permutations, the smallest p-value we can compute is $1/n$.

This approach is quite flexible and not limited to, say, the computation of t-statistics for each permutation. It is, however, easy to bias the results by not setting up the permutation procedure carefully. For example, in the above example, one might have wanted to directly permute the vector of 20 condition labels across all individuals.[13] These permutation would include many cases where both observations from an individual are assigned the same condition label and thus the difference statistic computed for each permutation would reflect a substantial amount of between-subject variance (i.e., this permutation test would not correspond to a paired *t*-test, because it does not take into account dependencies between multiple observations from the same individual).

So far we have described how to set up a permutation test for an individual hypothesis. For testing multiple hypotheses using the permutation test, one can apply the same procedures for controlling the FWER or the FDR to the *p*-values from a family of permutation tests. However, there is a way to incorporate control of the FWER directly into the permutation procedure. For each permutation one can construct the distribution of the *maximal statistic* (Nichols & Holmes, 2001) and then compare the statistic for the actually observed data against this distribution. For example, consider the example above, but instead of comparing two areas, we are interested in calculating the difference between the two ERP waveforms at each sample over a 1 s period sampled at 1000 Hz. For each permutation we can now calculate 1000 difference statistics (one for each sample) and retain the largest. We then compare all difference statistics for the actually observed data to this distribution of maximal statistics. Calculating the *p*-values by determining the proportion of more extreme statistics in this distribution of maximal statistics controls the FWER.

Again, it is easy to bias the results if one does not set up the permutation procedure carefully. For example, in the above use-case, one might be tempted to permute the condition labels separately for each sample across the 1 s period. In this case the difference statistic at each sample would be computed from independent assignments to condition labels and the distribution of maximal statistics would reflect between-subject variability in the auto-correlational structure of the time series.

Maris and Oostenveld (2007) have proposed additional refinements to the permutation procedure that capitalize on the auto-correlational structure in electrophysiological data, rather than just equate it for each permutation. The basic idea is that psychologically or electrophysiologically meaningful effects should be clustered (for example in time or space), whereas isolated effects are more likely due to noise. A difference in two ERP waveforms, for example, that is confined to a single sample (corresponding to 1 ms, for a sampling rate of 1000 Hz) is highly suspicious, because we would expect meaningful effects to last longer. This cluster-based permutation procedure requires the specification of the expected cluster size which needs to be done *a priori* (i.e., without being informed by the data that is being analyzed; see above). Naturally, if multiple cluster sizes are being considered, additional steps are required to control the FWER or the FDR as explained above.

*Bootstrap Methods**

A resampling method that is closely related to permutation tests is the bootstrap method (Efron & Tibshirani, 1993). Because permutation tests rely on the *exchangeability assumption* (i.e., condition labels can be exchanged under

[13] This would result in 20! (more than two quintillion) possible permutation and thus would be good example for a situation where it would be advisable to use a random sample of all possible permutations.

the null hypothesis), they test whether two distributions are identical and cannot test specific moments (such as means) in isolation. Usually this is not a problem—if an experimental manipulation truly has no effect, we would expect associated waveforms to be identical in all respects and not, for example, differ in variance even though the means are identical. However, it can be useful to estimate statistics of the data set, without relying on the *exchangeability assumption* or assumptions about the shape of the sampling distribution.

The bootstrap method consists of repeatedly sampling (with replacement) from each condition and calculating the statistic of interest based on these samples. Let us reconsider our above example with 10 participants for each of whom we have two ERP waveforms corresponding to distinct experimental conditions. If we repeatedly draw (with replacement!) 10 samples from the data set and calculate a t-statistic on the difference for each of these draws, we can calculate the bootstrapped p-value by comparing the t-statistic for the actual data with the distribution of bootstrapped t-statistics (the p-value corresponds to the proportion of bootstrapped t-statistics that are at least as extreme as the actual t-statistic). In each iteration we could also simply compute the differences; the standard deviation across these bootstrapped differences would correspond to the bootstrap estimate of the standard error of this difference.

The bootstrap estimate is asymptotically correct, in the sense that as the original sample size approaches the population size (which may be infinite) the bootstrap sampling distribution approaches the population sampling distribution. To deal with multiple comparisons, the bootstrap method can be also be used to estimate a maximal statistic or multiple bootstrapped p-values can be subjected to procedures to control the FWER or the FDR as explained above. Westfall and Young (1993) provide a comprehensive review and formal exposition of resampling methods for testing multiple hypotheses.