Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Comparison of Common Amplitude Metrics in Event-Related Potential Analysis

Karen Nielsen & Richard Gonzalez

MULTIVARIATE BEHAVIORAL RESEARCH
2020, VOL. 55, NO. 3, 478–493
https://doi.org/10.1080/00273171.2019.1654358

Comparison of Common Amplitude Metrics in Event-Related Potential Analysis

Karen Nielsen and Richard Gonzalez

University of Michigan

ABSTRACT

Waveform data resulting from time-intensive longitudinal designs require careful treatment. In particular, the statistical properties of summary metrics in this area are crucial. We draw on event-related potential (ERP) studies, a field with a relatively long history of collecting and analyzing such data, to illustrate our points. In particular, three summary measures for a component in the average ERP waveform feature prominently in the literature: the maximum (or peak amplitude), the average (or mean amplitude) and a combination (or adaptive mean). We discuss the methodological divide associated with these summary measures. Through both analytic work and simulation study, we explore the properties (e.g., Type I and Type II errors) of these competing metrics for assessing the amplitude of an ERP component across experimental conditions. The theoretical and simulation-based arguments in this article illustrate how design (e.g., number of trials per condition) and analytic (e.g., window location) choices affect the behavior of these amplitude summary measures in statistical tests and highlight the need for transparency in reporting the analytic steps taken. There is an increased need for analytic tools for waveform data. As new analytic methods are developed to address these time-intensive longitudinal data, careful treatment of the statistical properties of summary metrics used for null hypothesis testing is crucial.

KEYWORDS

Time series analysis; intensive longitudinal data; power; Type I error control; electroencephalography (EEG)

Introduction

This article focuses on a specific analytic protocol for commonly-used metrics in event-related potential (ERP) studies as an example of how a combination of analytic work and simulation study can uncover shortcomings of existing methods for time-intensive longitudinal data and drive the development of new methods. ERP studies are a prominent fixture in the psychophysiological literature. An ERP is a brain response to a time-locked stimulus or response. These brain responses are commonly measured using electroencephalography (EEG) scalp electrodes to capture voltage fluctuations at a high sampling rate while a research participant is engaged in a study task. Typically, the reported voltage is relative to another recording site on the scalp. Physiologically, changes in the voltage at the scalp are a result of the aggregation of many neurons firing (Buzsáki, Anastassiou, & Koch, 2012). For this reason, the underlying ERP waveform is anticipated to be a smooth progression of positive and negative voltage deflections. These shapes indicate underlying components, such as the P300 that denotes a positive deflection near 300 ms. Figure 1 (created using code from Helwig (2015)) shows the first 500 ms of a prototypical ERP waveform with three commonly studied components.

Given the anticipated hill shape of the components, a key element of ERP data analysis is the amplitude of each component. We focus exclusively on a popular technique described in detail by Luck (2014) and implemented in the MATLAB ERPLAB toolbox (Lopez-Calderon & Luck, 2014). The steps for this technique are outlined later in this article. Other approaches exist to answer slightly different research questions: for example, independent component analysis (ICA) decomposes the signal into several waveforms representing components (Makeig, Bell, Jung, & Sejnowski, 1996). It can be used to smooth data, such as for artifact correction. ICA is thus a more data-driven approach that allows for exploratoration, compared to the confirmatory technique detailed here.

ERP researchers use different methods to quantify and assess amplitude as well as for aggregating those amplitude metrics over trials. We begin by

CONTACT Karen Nielsen karenen@umich.edu Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48104, USA.

Code used for analyses, simulations, and creation of figures is available at: https://doi.org/10.5281/zenodo.3354464.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hmbr.

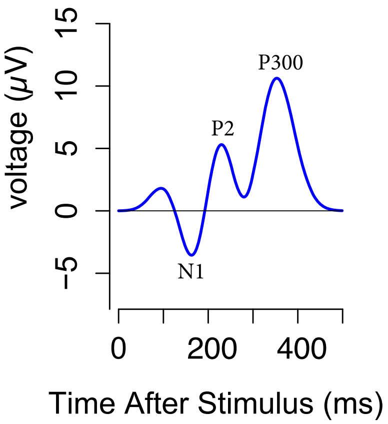


Figure 1. ERP waveform with three common components: N1, P2, and P300.

highlighting two summary metrics commonly used to quantify the amplitude of a component. One metric takes the average voltage in a prespecified window around a hypothesized component, while the other takes the peak, or maximum, voltage within a window. These summary measures can be considered special cases of a more general weighted average.

There does not appear to be an established standard for the definition of amplitude within the larger body of ERP researchers publishing in major journals. For example, of the 49 articles using ERP data published in the journal *Psychophysiology* in 2014 (some of which contained multiple studies and may be counted more than once here), 37 define amplitude as the average voltage in a prespecified window, 11 define amplitude as the maximum voltage in the window, and 5 use a combination such as the average near a local maximum, which is again a special case of a weighted average.

Several ERP analysis guides (Cohen, 2014; Dien & Santuzzi, 2004; Donchin & Heffley, 1978; Duncan et al., 2009; Keil et al., 2014; Luck, 2005, 2014; Murray, Brunet, & Michel, 2008; Picton et al., 2000) provide best practices for these summary metrics. Occasionally, recommendations are discussed in terms of potential impact on analysis (such as bias), but these guides typically do not include relevant theoretical background nor give much by way of boundary conditions of when to expect similarities and differences between these measures. Additionally, the recommendations do not cover the effects of the summary metric choice on statistical analyses such as Type I error rate and statistical power. In this article, we explore these aspects of each summary metric and focus on the common analytic approach used in the existing literature. To highlight the issues surrounding the choice of amplitude metric we do not consider other modeling details such as the error structure of the time series, nonparametric approaches, preprocessing and artifact detection issues, filtering, and data smoothing methods. We will also discuss the issue of aggregating curves across multiple trials and/or across multiple participants.

Several elements make it worthwhile to study the relative properties of these amplitude measures. Extreme value theory, a branch of statistics dealing with order statistics and extreme values, suggests concern about the use of a local maximum in a context where the average is otherwise appropriate. The maximum has been well-studied in the statistics literature (see, e.g., David & Nagaraja, 1970) and is known to have different distributional properties from the average, which has distributional properties that follow from the central limit theorem. Sometimes new measures, such as hybrids of the maximum and average, are proposed in the applied literature without sufficient theoretical justification nor evidence that the new measures exhibit desirable error rates and statistical power.

Clayson, Baldwin, and Larson (2013) review research comparing summary measures of amplitude as well as provide simulations under various noise conditions to evaluate their bias and variance properties. Their overall recommendation is that the maximum should be avoided in favor of an average. Their simulations, however, focus on estimation from a single subject fixed at 30 trials. The present article extends the work of Clayson et al. (2013) in several ways. In order to gain deeper understanding of the role of bias, variance, and other properties it is necessary to vary the number of trials. While Clayson et al. (2013) evaluates metrics by assessing their bias in parameter recovery, we focus on biases in null hypothesis testing. We also go beyond single-subject properties to make use of fundamental results in mathematical statistics to evaluate the differences and similarities between these estimators of amplitude. Further, analysis of ERP data frequently involves multiple subjects across two or more conditions so statistical tests such as analysis of variance (ANOVA) are conducted to test condition differences. We evaluate the performance of these amplitude summary measures with respect to their Type I error and statistical power across several settings. This approach reveals that there are cases where the maximum outperforms the average.

We begin by describing how ERP data are typically analyzed. We then give an illustrative example of a case when test results in the context of a paired *t*-test across subjects for the maximum and average do not agree. This example will serve as the motivation for the remaining sections, where we explore distributional theory and use simulations to determine how related decisions interact with the summary measure to impact results and interpretations. We conclude with an overview of the findings and discuss newer modeling approaches that may circumvent some of the issues uncovered in this article.

Analysis of ERP data

We first outline how ERPs are traditionally generated, collected, and analyzed. This outline has been adapted from several manuals, including ERPLAB (Lopez-Calderon & Luck, 2014), Cohen (2014), Keil et al. (2014), and Luck (2014), to reflect the standard accepted pipeline for ERP analysis. We will use this protocol throughout the remainder of this article.

1. Collect data in a continuous stream using EEG, coding the time of stimulus onset for each trial, the trial condition, electrode, and any other important events such as correct or incorrect response and response onset.
2. Clean data:
   a. Filter data to remove long-term trends or drift (for more information on filters, see Cook & Miller, 1992).
   b. Remove or correct artifacts (such as eye blinks, sneezes, coughs, etc.) using, for example, regression or ICA (see Keil et al., 2014, p. 6, for a full list and references).
   c. Re-reference data (optional). Standard references are a mastoid or the average of all sensors. The choice of reference electrode can be impactful (Dien, 1998), but can be leveraged to explore spatial relationships (Joyce & Rossion, 2005).
   d. Epoch the continuous data to create single-trial EEG segments (e.g., from −200 to +800 ms relative to stimulus onset).
   e. Baseline each trial so that reported voltages are relative to voltages prior to the key stimulus or response onset.
3. Average the single-trial EEG epochs (by taking the average at each time point) to create single-subject averaged ERP (AERP) waveforms for each condition of interest. Typically, each channel is treated separately. To avoid confusion with the averaging done in the next step, we refer to the waveform that results from this step as the AERP, in keeping with McGillem and Aunon (1987). Averaging is done here to gain a higher signal-to-noise ratio, by averaging out the variability of individual trials (Cohen, 2014).
   a. Some researchers take difference waves to compare two conditions (see, e.g., Kutas & Hillyard, 1980).
   b. Low-pass filters may be applied here. Low-pass filters attenuate high-frequency signals in order to increase the signal-to-noise ratio, in much the same way that averaging over trials smooths out the waveform.
4. Using a prespecified window to isolate the component of interest, calculate the summary measure (such as maximum or average) for each condition-level AERP for each individual. Thus, each subject is typically assigned a single-value summary measure for each condition-level AERP.
5. Perform an ANOVA for group-level analysis; this could be a between-subjects, within-subjects or mixed ANOVA depending on experimental design.
   a. Corrections such as Greenhouse-Geisser may be used for omnibus tests in the case of within-subjects analyses depending on assumptions made about the error covariance matrix (Jennings, 1987).

In this article, we focus on Steps 3, 4, and 5 to examine how use of the maximum or the average as a summary statistic of the AERP influences the statistical properties of hypothesis testing in Step 5. Thus, we extend the previous work of Clayson et al. (2013) by examining the effect on hypothesis testing and coverage probability. One of the most common tools for processing and analyzing ERP data is a MATLAB toolbox called ERPLAB (Lopez-Calderon & Luck, 2014). ERPLAB implements initial processing steps such as computing the AERP (Step 3) and summary measures (Step 4). Statistical analyses (Step 5) are completed in a separate statistical software of the user's choosing. Luck (2014) argues that the order of operations will not impact final results, so we also discuss whether the underlying distributions are invariant to reordering—particularly, of Steps 3 and 4.

Comparison of metrics: an example

To illustrate the potential differences in results based on either the maximum or the average, we examine a single component on a single channel in an ERP study. Reproducible code is available (Nielsen &

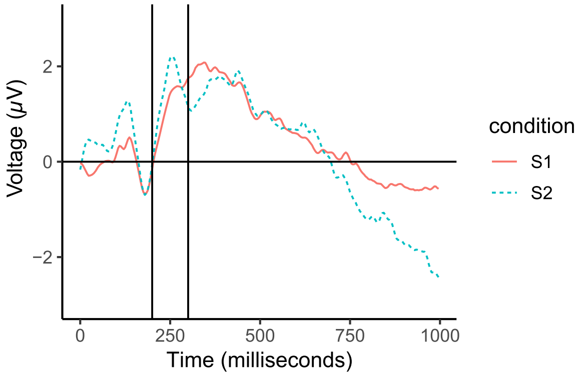


Figure 2. Grand average waveforms comparing experimental conditions ($N = 42$).

Gonzalez, 2019). We selected data from a publicly-available data set described by Zhang, Begleiter, Porjesz, and Litke (1997). The original study focused on ERP responses to paired images displayed in sequence that either matched or did not match. We retain the 42 control participants who provided complete ERP recordings following two task conditions—S1, in which the first picture is displayed on a computer screen, and S2 non-match (referred to here as simply "S2"), in which the second picture is displayed and does not match the first image. Time zero in each ERP recording corresponds to the beginning of the 300-ms image presentation and the next image does not appear until after the ERP recording ends. As a result of the study design, the S1 condition is more common (12–56 trials per person, mean = 44.0 trials) and S2 more rare (10–30 trials per person, mean = 21.83 trials). The unequal number of trials in the two conditions turns out to be critical as we will show. The data has been baselined and referenced to Cz after being recorded with a bandpass filter between 0.02 and 50 Hz, and we do not conduct any additional filtering. We assume that all trials containing artifacts have been removed in the publicly-available data. We elected to study a positive component near 250 ms using a window from 200 to 300 ms and the Cpz electrode for purely illustrative purposes.

A common presentation is to plot the AERP waveforms averaged across subjects, so that the two conditions can be compared qualitatively. Figure 2 shows the responses to the two conditions. To create this figure we followed standard procedure and first averaged each timepoint over all trials for each condition within person, and then averaged over people (to produce a "grand average" waveform).

Figure 2 shows differences in the two conditions over the majority of the one-second average recording, beginning at stimulus onset. In particular, the common S1 condition has, on average, lower voltage throughout the area of interest (near 250 ms). It is difficult to tell if this difference is significant or not because this type of plot has no representation of the underlying trial-to-trial and across-subjects variabilities. Instead, the plot shows the stability of the averaged process over time. The component of interest, a positive deflection centered near 250 ms, is visible in Figure 2. There is a potentially overlapping component in the S1 condition, and possibly the S2 condition as well, that may be contaminating the signal, particularly on the right side of the window.

Table 1. Comparison of maximum and average amplitudes from grand-averaged waveforms ($N = 42$).

	Maximum	Average
S1 (common) condition	1.7149	1.1419
S2 non-match (rare) condition	2.2109	1.4883

Table 1 shows the values for the maximum and average voltages for each condition in our window of interest (200–300 ms post-stimulus) based on grand average AERPs in Figure 2 and highlights two important details. First, the maximum is greater than the average within each condition. This will always be the case because mathematically the maximum is as large or larger than the average in any set of values. The table also confirms what is visible in Figure 2: in the window of interest, the rarer S2 condition has larger amplitude than the S1 condition as measured by both metrics. This table, like Figure 2, does not provide information about the trial-to-trial and across-subjects variabilities in the underlying data.

To test if the difference is statistically significant, we can perform a paired *t*-test using the summary measures. We have 2 summary measures for each of 42 participants, and thus can measure across-subjects variability of the differences controlling for intrasubject association between the two conditions and perform a significance test. However, we lose the ability to isolate trial-to-trial intrasubject variability due to the AERP in Step 3. Table 2 shows the results of the 2 paired *t*-tests for the maximum and average.

We can see that the maximum and average yield different *p*-values and significance test results at $\alpha = .05$. However, we do not know which test is correct. What are some of the contributing factors that might lead to the disparities in these statistical tests?

First, not only will the metrics take different values, but when viewed as random quantities, they will have different underlying distributions. To investigate this, we derive properties of the underlying distributions and pairwise differences of these quantities using reasonable assumptions for the ERP context. The

Table 2. Comparison of *t*-tests for maximum and average across conditions.

Test	Results
Paired *t*-test for maximum	$t(41) = -3.013$, $d = -0.831$, $p = 0.004$
Paired *t*-test for average	$t(41) = -1.501$, $d = -0.346$, $p = 0.141$

distributions of the two summary measures may also be differentially influenced by noise. In the across-subjects grand average AERPs in Figure 2, there are a few small spikes that create local maxima outside the window of interest. These are much more frequent, including inside the window of interest, and dramatic in the within-person waveforms, but they are attenuated by the additional averaging step used for plotting. The spikes, artifacts from the variability of each trial, allow for occasional large values for the maximum. This can have a major impact on statistical tests. To obtain a large difference within person for the maxima, only one large spike in a condition's AERP is needed, whereas a large difference in the average requires a more systematic difference in the two conditions throughout the window of interest. Filtering may reduce such spikes. We investigate the impact of noise on the performance of these summary measures.

Second, there may be latency differences in the timing of the component of interest at the individual level. The procedure for analyzing ERP amplitudes implicitly assumes that no latency difference exists between conditions. A latency difference might mean that the window location is not optimized for both metrics. A small miscalculation on the window location may not have a major effect on the maximum (or adaptive mean, discussed later) because as long as the peak of the component is in the window, these measures will be unaffected. The average, however, will change with small adjustments of the window. For example, if the component is symmetric and hill-shaped, we would want the window to be centered on the peak to find the largest value for the average. We can infer from Figure 2 that moving the window to the right would result in larger values for the average in the S1 condition, but perhaps not in the S2 condition. Also, a smaller window will yield larger values for the average in this case, but not for the maximum. Luck (2014) makes different recommendations for window size based on the chosen summary metric. He suggests using larger windows when assessing the maximum than when assessing the average due to the increasing risk of missing the peak amplitude with smaller windows, with a minimum window size of 50 ms for the average to ensure that high-frequency noise is attenuated. We will explore and discuss this rationale in more detail later in the article.

Third, the unequal trial counts in the two conditions may have influenced the relative performance and efficiency of these two summary measures. We know that unequal sample sizes in comparison groups can lead to misleading results if standard error calculations are not weighted properly, and small comparison groups can lead to the statistical test being underpowered. In the current example, trial counts are on average 44 per participant for the common S1 condition and less than 22 for the rarer S2 condition. Because we compute each subject's AERP to remove noise, the unequal trial count may be leading to unequal variances in the summary measures that ultimately impact test results.

Fourth, while the single observation nearest the time when the true maximum occurs may have negative noise, the dense sampling that is standard for ERP research will lead to several observations being collected near the true maximum; we can expect one or more of these to have positive noise. The value near the maximum with positive noise will be identified as the sample maximum—thus, the sample maximum can be expected to be positively biased. The average value, however, will not be subject to this bias.

We next review some relevant distributional theory and then discuss the potential factors listed above using both theory and simulation results.

Distributional theory of summary metrics

We use statistical theory to explore the asymptotic distributions and convergence rates of the summary metrics as they relate to ERP null hypothesis testing. In most ERP studies, ANOVAs are used to compare groups or conditions of interest on summary measures for each subject—such as the average or maximum. To simplify our analyses, we will focus on the simple two-condition version of repeated-measures ANOVA, the paired *t*-test. These tests carry several assumptions. For example, the differences are assumed to be normally distributed. It is not intuitively obvious if this assumption is met when testing maxima in this context and distribution theory can provide insight. We also consider how variance differences between summary metrics come about and how they will impact statistical tests.

One assumption is necessary to simplify the exploration of underlying distributions. Throughout this section, we assume that observations in a given time window are independent and identically distributed (IID). While observations of an EEG recording are clearly not temporally independent, the maximum and average do not make use of the temporal ordering of the data so we proceed under the assumption of independence. We assume that each time point has noise that is produced in a consistent way, as a combination of human physiology and the EEG equipment. Thus the observations are treated as identically distributed. Because to compute the AERP we first average at each time point over all trials for a condition (we further assume that the sampling rate is exact to simplify this step), we invoke the central limit theorem so the original distribution of the individual time points does not matter. Under the central limit theorem, each point along the AERP is normally distributed with variance proportional to the amount of noise present in the data. Further, the difference of two normals is also distributed normal, a convenient property when constructing tests of differences between means.

In ERP studies, there are several sample size decisions—each of which may impact results on top of the choice of summary metric. The number of data points m retained for study in a single trial is determined by the window size and the sampling rate, as described in Equation (1).

$$m = \lfloor(\text{window duration in seconds})* (\text{sampling rate in Hz})\rfloor \tag{1}$$

For example, if our window is two-tenths of a second and we use a 256 Hz sampling rate, $m = 51$. Further, the sample size contributing to each point along the AERP is dependent on the number of trials, n. These seemingly small choices can have substantial impact downstream, when we are depending on asymptotic results.

The distributional theory for the average is simple. The central limit theorem shows that the sum, and thus the average, of m IID observations tends toward a normal distribution as m tends to infinity. The rate of convergence is generally stated to be $\frac{1}{\sqrt{m}}$, and if the original distribution is already close to normal (as should be the case, because we first average over trials), only a small m is needed to make statistical tests with a normal distributional assumption reasonable. These rules of thumb can be explored more thoroughly by using the Berry–Esseen theorem.

The Berry–Esseen theorem helps us to estimate the sample size m needed for reasonable convergence when using the central limit theorem. The distance between F, the cumulative density function (CDF) of the IID samples in the AERP, and a normal distribution is bounded:

$$D \leq \frac{C\rho}{\sigma^3\sqrt{m}}, \tag{2}$$

where ρ is $E[|X_1|^3]$, the third absolute moment, σ is the standard deviation of F, and C is a constant less than .56 (Shevtsova, 2011) and greater than .41 (Esseen, 1956). The theorem gives an upper bound on the distance, i.e., it gives the worst-case scenario of the distance from normality. The two main elements that affect this upper bound are the sampling rate, m, and the ratio ρ/σ^3, which is impacted by the symmetry and variability of the distribution of observations in the AERP window of interest.

Asymptotic theory shows that averages within a window will approach a normal distribution that meets the distributional assumptions of t-tests and ANOVAs. Equation (3) describes the asymptotic result for the expected value of the average summary measure from the central limit theorem.

$$\sum_{i=1}^{m}\frac{X_i}{m} \sim N\left(\sum_{i=1}^{m}\frac{x_i}{m}, \frac{s^2}{m}\right) \tag{3}$$

Here, s describes the standard error of the AERP, which will depend on the inter-trial variability, the intra-trial variability, and the number of trials. Thus, the variance of the average summary measure depends on within and between trial variability. This walk-through of the distributional theory shows how the average summary measure is impacted by various elements of the study design and ERP signal.

The distributional theory for the maximum is not as simple. Just as the central limit theorem is used to characterize the asymptotic distribution of sample means, the Fisher–Tippett–Gnedenko theorem (also known as the extreme value theorem) provides the asymptotic distribution of the sample maximum. The Fisher–Tippett–Gnedenko theorem states that, if the distribution of the maximum converges, it must converge to one of three distributions: Gumbel, Frechet, or Weibull (Fisher & Tippett, 1928; Gnedenko, 1943). These three distributions have been generalized as special cases of the generalized extreme value (GEV) distribution. The theorem does not guarantee convergence or give criteria for convergence, but we will assume convergence for simplicity. Because we first compute the AERP over trials in the ERP context, each time point in the window should asymptotically have a normal distribution. Gnedenko (1943) showed that the maximum of a series of IID standard normal distributions follows a Gumbel distribution, which has a PDF given by

$$\frac{1}{\beta}e^{-\left(\frac{x-\mu}{\beta}\ +\ e^{-\frac{x-\mu}{\beta}}\right)}$$

with μ denoting the location parameter and β is proportional to the standard deviation. Both parameters can be expected to vary based on characteristics of the subject-level AERP and specification of the window of interest.

If we want to compare the maximum M_{m1} from one condition with the maximum M_{m2} of another condition, and we assume that the scale parameter (β) is the same for each underlying distribution, we can use the result that the difference of two Gumbel-distributed random variables with the same variance follows a logistic distribution. In other words, when we test that the location parameters for the two trials are equal, the null distribution of $M_{m1}-M_{m2}$ follows logistic$(0, \beta)$ (Arnold, 1992; Gumbel, 1958).

The logistic and normal distributions are similar (Balakrishnan, 2013) but not identical. The logistic distribution is slightly more peaked and has slightly wider tails than the normal distribution (Chew, 1968). While the Berry–Esseen theorem gives convergence rates to normality, there is not equivalent theory to suggest a convergence rate for the convergence to Gumbel (and thus, of the differences to logistic). However, we can already see that many of the same elements will impact the maximum as did the average, but in different ways.

The order of operations for simplifying ERPs is important. If our methodology involved taking the summary measure in each trial before averaging over all trials in each condition for each person, then the central limit theorem, the Berry–Esseen theorem, and normality would all hold regardless of the summary measure and we might expect thedownstream test results to be similar. Under the current protocol, only the average is unaffected by reordering Steps 3 and 4, assuming equal sample sizes or proper weighting of observations in an unbalanced design.

Simulating data from an ERP component

We now turn to a series of simulations to explore various design and analysis decisions that were raised when reviewing the standard analytic steps in ERP research and potential explanations for the different inferential results between the mean and the maximum in the example paired t-tests presented earlier in the article. Reproducible code for all simulations and figures in this section is available (Nielsen & Gonzalez, 2019). This simulation allows for careful exploration of single aspects of ERP studies because the true results are known and all other aspects are held fixed. In addition to comparing the mean and the maximum we also include the hybrid measure briefly mentioned in the introduction. This hybrid method, sometimes referred to as the adaptive mean, was investigated alongside the maximum and average by Clayson et al. (2013). The hybrid measure identifies the maximum amplitude in the window for the condition, then takes an average of the 5 points surrounding and including this peak. Only points in the window are used. Because of the way it is constructed, the hybrid measure should converge to normal or logistic faster than the maximum converges, but not as fast as the average does, so one might expect it to perform between the two measures. Indeed, Clayson et al. (2013) found this to be the case. However, some users seem to think that it will outperform both because it seemingly avoids the known pitfalls of both.

The simulation focuses on only one ERP component, for one channel (which we assume to be recorded relative to a reference). We make the assumption that our simulated data reflect ERP data that have been acceptably filtered for the protocol described in this article—that is, as much irrelevant noise as possible has been removed. To achieve a hill-shaped component, we use a normal kernel. The choice of normal kernel has been used by Helwig (2015) as a way to recreate visual-stimuli ERP waveforms based on data from several studies. Helwig's eegsim function in the eegkit package in R uses a prespecified voltage weight for each channel and multiplies the functional form of a normal kernel by this weight to simulate ERP components. We generalize this process by defining the voltage v_i at each time (in seconds post-stimulus) t_i as:

$$v_i = He^{-W(t_i-L)^2} \tag{4}$$

where H determines the height of the component, W determines the width, and L is the latency of the peak of the component relative to stimulus presentation. This shape returns to 0 on either side of the hill-shaped component. Thus we can assume that the baselining step has already been performed and there are no overlapping components.

We account for variability in the timing of ERP components and allow for individual differences for each subject j in terms of both the height of the waveform and the latency. Trial variability for each trial k is modeled similarly. Thus, voltages are generated as follows for each trial in the study:

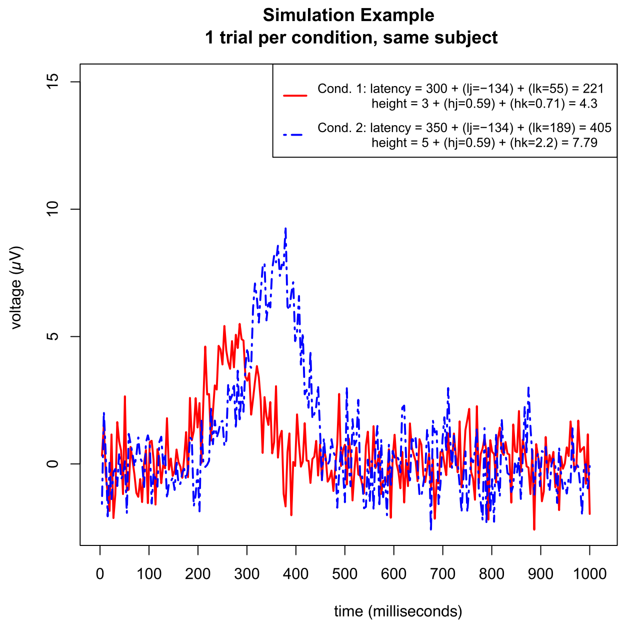


Figure 3. Example of simulated data from 1 trial per condition.

$$v_{ijk} = (H + h_j + h_k)e^{-W(t_i-(L+l_j+l_k))^2} + z_i.$$

This formulation is consistent with a random effects model allowing for heterogeneity over trial k and subject j in the height and latency. The model also includes random noise at each time point of each trial, z_i. We assume that noise at a specific time point is what remains after filtering and is a function of the equipment and overall human physiology rather than being related to the subject or condition.

This simulation design allows for variation at each of the levels of the naturally-occurring hierarchy, but has been designed to exclude interactions across these levels. For example, as described below, the distribution of trial-level height and latency h_k and l_k are not related to the person j.

We focus on Type I and Type II error rates for a study in which 20 participants experience two conditions that produce effects on the component being studied, such as a go/no-go task. This sample size was selected to reflect typical ERP studies (e.g., the midpoint of the sample sizes in the meta analysis by Umbricht and Krljes (2005)). Type I errors, or false positives, can lead to incorrect claims of differences between groups or conditions. To investigate the Type I error rate, we simulate 10,000 studies, in which all trials are generated using the same values, but are labeled as belonging to two separate conditions. We then follow the testing protocol presented earlier and report the proportion of times the two conditions yield significantly different results ($p \leq .05$) for each summary measure. The simulation sets the known population values as follows. The true height H is set at 5 to reflect a peak amplitude of 5 microvolts. This value sets the scale of the vertical axis, and is thus arbitrary. W is fixed throughout the study at 200 (i.e., a component width of approximately 200 ms) because the width of the component is not generally of interest in ERP studies. The latency, or time of peak amplitude, L is set to .300 for a component that is centered at 300 ms. Thus, all voltages are generated from the following equation for Type I error assessment:

$$v_{ijk} = (5 + h_j + h_k)e^{-200(t_i-(.300+l_j+l_k))^2} + z_i. \tag{5}$$

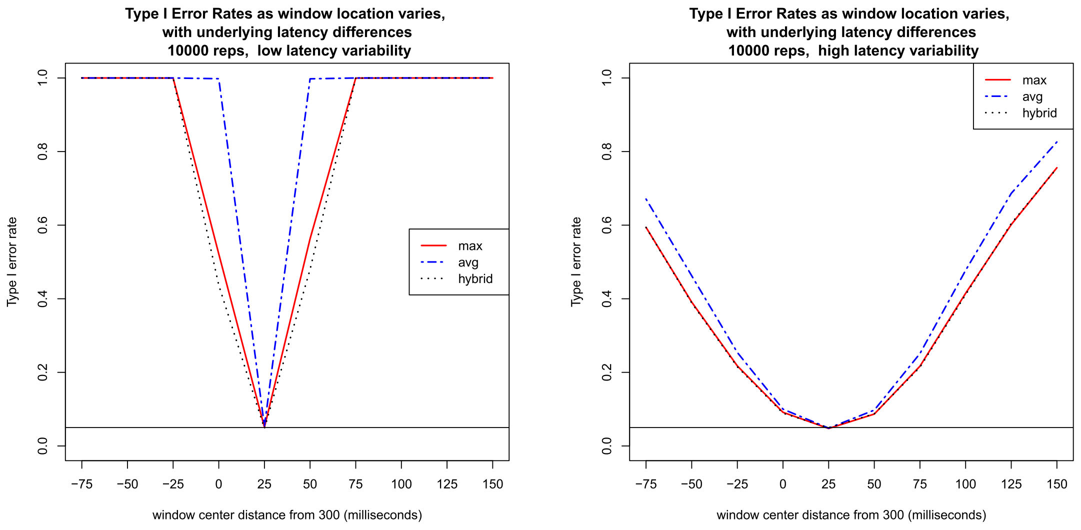
Polich and Kok (1995) explore many sources of P300 amplitude and latency variability, at both the trial and subject level. Our parameter values throughout this simulation are selected to generate data with variability within the range of these effects that would often be averaged over in an ERP study. Specifically, subject and trial height adjustments h_j and $h_k \sim N(0,1)$ reflect amplitude variability at each level, at a scale appropriate relative to H. The noise at each time point, z_i, also impacts the recorded amplitude. Throughout the simulations, $z_i \sim N(0,1)$ unless otherwise stated. We explored smaller values for the variance of z_i, which resulted in smoother waveforms at the trial level, and found only one case where this parameter impacted error rates. Because low-pass filtering data is typically done to reduce noise and smooth waveforms (Luck, 2014), we anticipate similar findings would hold if we were to conduct a similar study of low-pass filters. Unless otherwise stated, l_j and $l_k \sim N(0, 0.100)$ to reflect the substantial variability in latency at the subject and trial levels. We explore the impact of this latency variability on amplitude estimates at several points throughout the simulation study.

We use a similar approach to investigate Type II errors. Trials labeled as condition 1 are generated from Equation (5). A second condition, in which the true population amplitude is lower (3 μV instead of 5) and the peak latency is slightly later (350 ms instead of 300, again following Polich & Kok, 1995), is described in Equation (6). The subject and trial level variables are specified as before.

$$v_{ijk} = (3 + h_j + h_k)e^{-200(t_i-(.350+l_j+l_k))^2} + z_i. \tag{6}$$

The Type II error rate reflects the proportion of times a true difference fails to be found (i.e., $p > .05$) using the given experimental design.

There are additional variables manipulated in the simulation that reflect aspects of the study design that



(a) sd = .020

(b) sd = .100

Figure 4. Type I error rates as window centering varies, conditions differ on component location but not amplitude.

are under the control of the researcher. For example, the sampling rate (measured in Hz, or observations per second) will determine the time points t_i at which data are available. We assume that the time points are equally spaced, start at stimulus onset, and extend to one second to produce m samples; that is, if we simulated at 4 Hz, we would have $m = 5$ samples at $t_i = \{0, .25, .5, .75, 1\}$. The sampling rate is 256 Hz unless it is specifically being investigated. We also vary the number of trials per condition, n. This value is fixed at 30 when not otherwise stated. The remaining element that is varied is the size and location of the window used to calculate the summary statistic (either maximum or average). A window from 250 to 350 ms is the default. The number of subjects in the study is assumed to be fixed at 20 throughout. We assume for simplicity that no trials are dropped from the study.

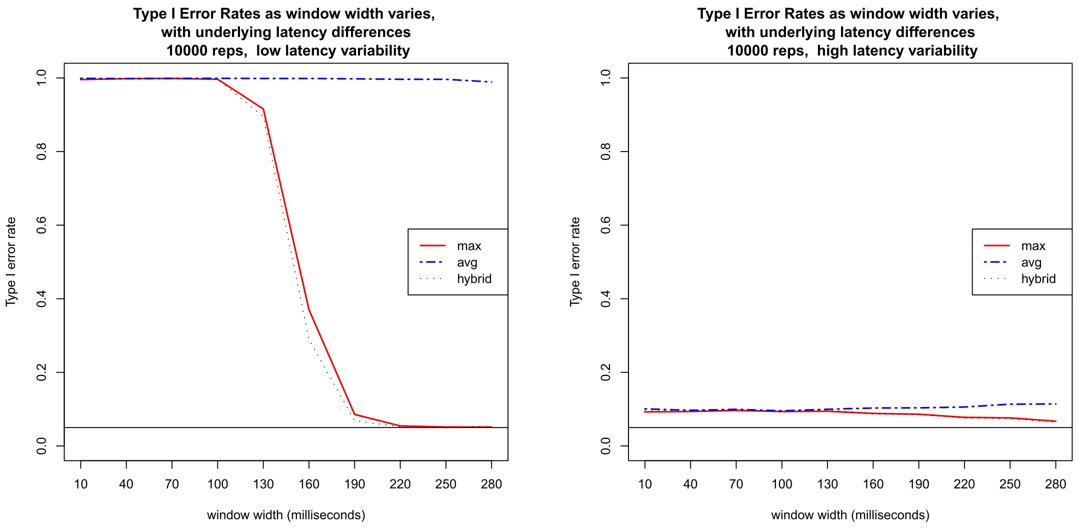
Figure 3 shows examples of two simulated trials from the same participant for different conditions. The stated parameter values are sampled randomly from the distributions described above to produce underlying height and latency values for each trial, and the sampling rate is 256 Hz. Each timepoint is subject to IID noise, $z_i \sim N(0, 1)$.

The results of the simulations are organized into three subsections. The first subsection focuses on the impact of latency across window locations and widths, the second subsection examines the role of noise and the number of trials on Type I and Type II error rates, and the third subsection focuses on the role of the sampling rate on Type I and Type II error rates. We also used this simulation framework to conduct checks on the distributional properties, such as whether the maximum follows the Gumbel distribution and whether differences between two peak amplitude measures are approximately normally distributed (i.e., whether the normal does a reasonable job of approximating the logistic). These simulation results support the underlying distributional theory so will not be presented here.

The impact of latency across window locations and widths

A latency difference in the timing of the component of interest between conditions may impact results. In this section, we start by exploring the impact of variability in latency—both between-subjects and trial-to-trial (i.e., within-subjects)—over a range of window specifications. We then consider the case with actual differences in the true latency L and how this impacts tests of the amplitude. We assume that only the amplitude is of interest and explore how error rates differ depending on the researcher's specifications for window width and location.

Relatively little research has examined issues of window specification and latency jointly though some articles examine them separately. For example, Luck (2014) recommends a narrower window when working with the average than the maximum, with a width of around 50 ms for the average. Picton et al. (2000) emphasize that increased variability in latencies will lead to smaller amplitude estimates due to the averaging step. However, we have not seen these elements of an ERP study discussed together.



(a) sd = .020 (b) sd = .100

Figure 5. Type I error rates as window width varies, window centered at 300 ms, conditions differ on component location but not amplitude.

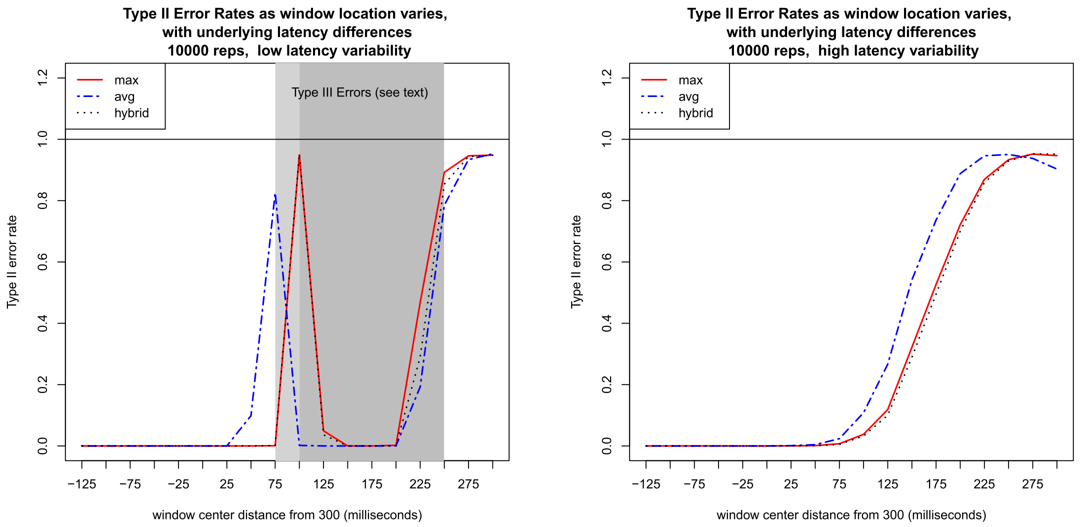
We start by examining Type I error rates across different window sizes, then window locations, comparing these results for two possible variances defined by the variables l_j and l_k. Using Equation (5), we simulate 10,000 studies in which 20 participants experience 30 trials each in two conditions where the null hypothesis is true—no differences exist in either the amplitude or the latency. We find that the choice of window width and location does not matter. The Type I error rate remains near .05 and the Type II error rate near 0, whether the variability in latencies is high (sd = .100) or low (sd = .020). This is promising because so far we have only added noise to a variable that is not of interest.

Next, we introduce signal by considering the case where amplitudes between the two conditions do not differ, but latencies do. In this simulation, $H = 5$ for both conditions, but $L = .300$ for condition 1 and $L = .350$ for condition 2; that is, condition 2's true peak amplitude occurs at 350 ms instead of 300. Because the stated hypotheses and tests concern only amplitude, and component amplitude does not differ across the two conditions, a significant difference is considered a Type I error as it would be incorrectly reported as an amplitude difference. We hold the window width fixed.

We find that if there is a true latency difference between the two conditions, the window must be centered near the midpoint of the two condition peaks to minimize Type I error rate. Otherwise, the error rate becomes inflated at a rate relative to the latency variability. Figure 4(a) shows that in the case where latency variability is relatively low (i.e., $l_j, l_k \sim N(0, .020)$ and a true latency difference between conditions of 50 ms), deviations from this ideal can lead to a rapid inflation of the Type I error rate. This is likely due to the high signal-to-noise ratio on the latency difference.

Figure 4(b) shows that when the latency variability is higher (i.e., $l_j, l_k \sim N(0, .100)$ and a true latency difference between conditions of 50 ms), deviations from the centering ideal again lead to inflation of Type I error rates, but at a slower rate than when the latency variability is low.

Regardless of the signal-to-noise ratio, the maximum and the hybrid measures are more robust to errors in window location because any window that contains both of the true peaks, even if it does not contain the entirety of both components, will allow for the correct testing of the amplitude difference when using these summary measures. This also means that a wide enough window can overcome a non-ideal window location for these two measures, but not for the average. To demonstrate this, we use the same setup to explore results as the window width expands from a center at 300 ms. Figure 5 shows how larger window widths interact with this non-ideal choice for the window center. In Figure 5(a), the maximum and hybrid maintain Type I error rates near 5% for windows larger than 180 ms because the window is large enough to cover both components consistently due to the small between-subjects latency variability. However, when the window is smaller, these maximum-based measures miss the later condition's peak.



(a) sd = .020

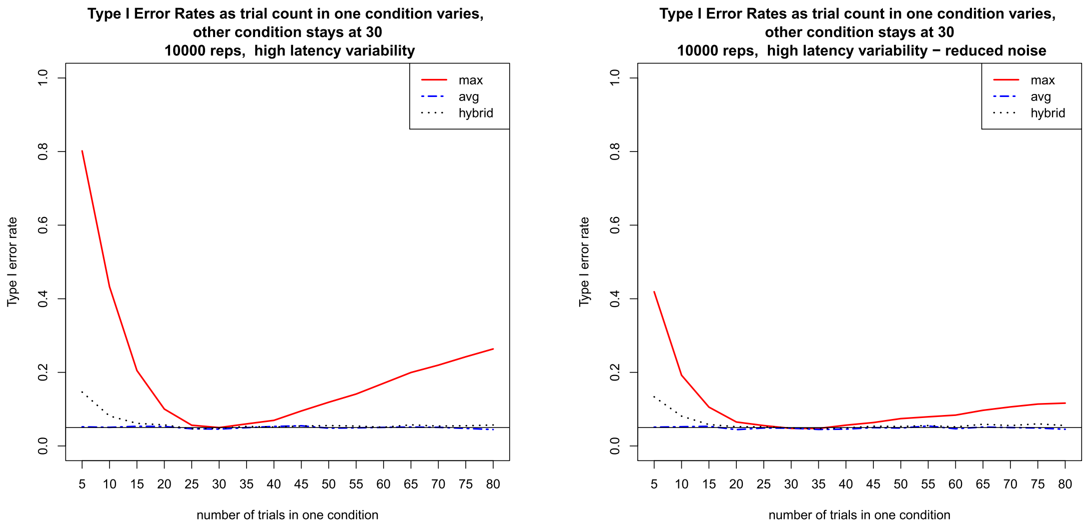
(b) sd = .100

Figure 6. Type II error rates as window location varies.

The average has a very high error rate, as seen in Figure 4(a) at 300 ms, and a larger window cannot help it recover. In Figure 5(b), the error rate remains near 10% for all three measures because the high latency variability leads to components regularly falling outside the range of a window of any studied size centered at 300 ms (note the 10% error rate at 300 ms in Figure 4(b)). In this case of high latency variability, the maximum and hybrid again outperform the average in large windows, but by a smaller margin.

We next inspect the Type II error rates by making use of Equation (5) for condition 1 and Equation (6) for condition 2 so that a component with lesser amplitude occurs at 350 ms and one with greater amplitude occurs at 300 ms. We now investigate how window location and latency differences may hinder discovery of this true amplitude difference. Figure 6 reveals additional trends in errors that occur when the window is not centered at the ideal location. As the window center is moved to later times, the Type II error rate rises. This happens because the earlier component is being cut off, producing smaller values for the summary measures than it should and making the two components seem to have similar amplitudes. As with the Type I error rate, the average is most sensitive to deviations from ideal specifications, and studies with smaller between-subjects variability see more rapid inflations of the error rates due to high signal-to-noise ratio in both the variable of interest (amplitude) and the variable not of interest (latency). Again, the hybrid measure tracks almost exactly with the maximum. The temporary recovery of the Type II error rate in Figure 6(a) is due to the 2-sided testing procedure beginning to find that condition 2 (the later component with lower amplitude) has greater amplitude than condition 1. Because we know the true data-generating models in this simulation, we know that these are incorrect findings and have labeled this region as Type III error. However, in a real ERP study using ANOVA, this could easily be overlooked and reported as the opposite result. This underscores the importance of careful inspection of the grand-averaged waveforms, proper corrections in post-hoc testing in exploratory research, and a solid hypothesis about the location of the component for confirmatory research. Figure 6(b) again highlights that high variability, at least in the variable not of interest, can be beneficial in this testing setup. Here, the components in individual trials fall within the window often enough for the amplitude differences to be detected over a greater range of window locations. Type II error rates remain near 0 for the range of window widths centered at 300 ms post-stimulus.

To summarize this subsection: the window size and location do not matter if the difference between the two conditions is only amplitude. However, if there is a true latency difference across conditions, these choices can be critical. Error rates are higher when the variability in latency is lower, and maximum-based measures are more robust to these choices. These errors could be avoided by using different windows for each condition if there is a known latency difference, but this difference may be difficult to identify in practice.



(a) $z_i \sim N(0,1)$

(b) $z_i \sim N(0,.5)$

Figure 7. Type I error rates in unbalanced designs.

Unbalanced designs and trial counts

In ERP research, studies with unbalanced designs such as those from oddball paradigms like that of Friedman, Cycowicz, and Gaeta (2001) are common. However, several articles (Luck, 2010; Thomas, Grice, Najm-Briscoe, & Williams Miller, 2010) have highlighted the dangers of working with unequal trial counts in ERP studies. In this subsection, we use simulations to explore the impact of the number of trials per condition on test results.

As before, we use Equation (5) to simulate 10,000 studies in which 20 participants experience two equivalent conditions, but here we hold the trial count fixed at 30 for one condition and allow it to vary for the other condition. Figure 7(a) shows that unbalanced designs can have dramatic impact. The maximum finds many more false positives when the number of trials in the two conditions is not similar, and even the hybrid measure has increased Type I errors in the most extreme cases. When we reduce the latency variability across individuals, a similar trend emerges. However, the error rates for the maximum are slightly higher when the latency variability is reduced. Perhaps lower trial-to-trial variability in the component's location can lead the individual waveforms in the condition with fewer trials to combine to a more-pronounced peak in that condition (for a visual explanation, see Figure 4.2 in Luck, 2014), leading to a higher amplitude and thus a significant difference.

Concerns about unequal trial counts in ERP research are not new. However, a single explanation has not been established. Polich (1986) and Luck (2014) argue that as the number of trials increases, the signal-to-noise ratio increases and the waveform has fewer distractor peaks as a result of the larger number of trials being averaged at each time point. These additional peaks that occur with smaller trial counts provide more opportunities to find a difference in the maxima, even when one does not exist. Thomas et al. (2010) give an alternative explanation in which the averaging process of Step 3 leads to decreased probability of obtaining an extreme value at the true peak due to noise as the number of trials increases.

An additional factor that impacts the signal-to-noise ratio is the extent of noise in the amplitude of collected individual timepoints. Figure 7(b) shows the impact of reducing the variance in noise at each time point by generating $z_i \sim N(0, 0.5)$ rather than $z_i \sim N(0, 1)$. By reducing the noise, we reduce the scale of the heterogeneity of variance that results from averaging unequal trial counts per condition in Step 3 of the procedure outlined earlier in this article. Varying the degree of noise in the simulated data may serve as a proxy for varying the extent of filtering intended to remove high-frequency noise, though it is important to note that filters often have other unintended side effects (Luck, 2014).

All three metrics yield low Type II error rates in studies with unbalanced trial counts across conditions, regardless of the noise z_i. This remains true with lower latency variability. This matches the argument by Luck (2010) that reducing the trials in one condition will not increase the Type II error rate. There were a few errors where the average failed to yield significant differences when the trial count in one condition was particularly low, but even these errors were rare.

The simulations in this section suggest that the maximum may underperform in studies with unbalanced designs, particularly when no true difference exists and data are noisy at the trial level. The hybrid is more robust, but still has an elevated Type I error rate in extreme cases. While smaller sample sizes are generally associated with increased error rates and results that do not replicate (Button et al., 2013), the Type I errors we highlight here occur any time the trial counts in two conditions are unbalanced. Several possible solutions to this problem have been proposed. Trials can be thrown out at random from the condition with a larger trial count (Clayson & Miller, 2017). However, this results in a loss of data. Instead, the average could be used exclusively when dealing with unequal sample sizes (Thomas et al., 2010), or a random effect approach could be used (described in the discussion). Reducing noise, perhaps via low-pass filtering, may greatly reduce these Type I errors.

Sampling rates

Another design feature that may affect results is the sampling rate. Equation (1) describes how the number of observations being used to calculate the summary measure is determined by both the window size and sampling rate. As with the previous subsections, we simulated 10,000 studies where both conditions were generated from Equation (5). We repeated this procedure over a range of feasible sampling rates for EEG research.

The Type I error rate is stable at .05, with no differences between the three summary metrics, as sampling rate varies from relatively low (100 Hz) to much higher than is standard for ERP studies (1000 Hz). Additionally, in our simulation no Type II errors occur, regardless of summary metric. This suggests that our simulation was well-powered. These results hold for high or low latency variability. Some authors propose that the choice of ERP sampling rate should be determined by the Nyquist rate (treated in more detail in Proakis & Manolakis, 1992)—the sampling rate should be at least double the rate of the phenomena of interest. Because ERP components tend to last on the order of one tenth of a second, a sampling rate of at least 20 Hz should be adequate. However, there may be high-frequency noise that needs to be filtered out, and so the sampling rate should be selected in order to capture, then filter out, this noise. We do not explore filters in this study, but do not anticipate that they will solve the inflated error rates we have uncovered elsewhere in our simulations—instead, they may introduce additional complexities to the analytic decision-making process.

Throughout these simulations, we explored properties of the average, maximum, and hybrid summary metrics. We highlighted the importance of the window width and location when there is an unmodeled latency difference across conditions. We revealed that unbalanced trial counts lead to inflated Type I error rates for the maximum and hybrid, but not the average. We found that sampling rate does not appear to impact results for these metrics, as long as sampling rates are within the range of typical EEG equipment.

Discussion and future work

We explored the use of three common summary metrics in the context of ERP analysis. The motivating example demonstrated that summary metrics can yield contradictory results for statistical tests. We reviewed theoretical findings, such as extreme value theory, that become relevant when working with different summary measures like the maximum. We then used simulations to explore the differences in Type I and Type II error rates across several common experimental scenarios. These simulations revealed a few valuable findings about the performance of these summary metrics.

Our first major finding is that the choice of window location is critically important when the conditions have unmodeled latency differences—that is, when the analyses are concerned only with peak amplitude, but there is a systematic difference in the timing of the component across groups or experimental conditions. Type I and Type II error rates increase as the window moves away from ideal placement, and the error rates increase more rapidly when there is smaller variability in latency across subjects. The maximum and hybrid are more robust to window choice, and increasing the width of the window can also help to capture the peak amplitude. These findings underscore the need for researchers to take steps to confirm that their test results are not driven by unanticipated factors. For example, we recommend that researchers check visualizations for this potential confounder.

A second major finding is that the use of the maximum is more likely to lead to a false positive finding when comparing groups in unbalanced designs, such as oddball paradigms. We believe this is due to the unequal amplitude variances that result from first averaging across trials within person. Error rates are lower (but still higher than nominal) in cases of reduced noise. Thus, the recommendation we make is to use averages when analyzing such studies, and to be aware of the potential for false positives when evaluating studies with unbalanced designs that make use of the maximum. Filtering and smoothing procedures may prove beneficial in this situation.

Overall, we find that each summary metric has strengths and weaknesses. The hybrid is not, in the situations we explored, a cure-all method. Anticipated features in the data, the costliness of particular errors, or the desired interpretation may need to be stated to justify the choice of summary metric in the current analytic paradigm. This conclusion is frustratingly nonprescriptive, particularly for a problem where consistent methodology recommendations may be needed for replicating studies. But when one uses multiple metrics such as fit to theoretical distribution, Type I error rate, and Type II error rates, one procedure may not always dominate across all conditions for all metrics that one considers. Further, we note that the Type I error rates are at times quite discrepant from their nominal values, reaching extreme Type I error rates of .90 or greater. These findings underscore the need for detailed methods sections, describing and justifying all choices made between data collection and final test results. Researchers should also consider making their data available to others, who may wish to explore the impact of alternative analytic protocols.

The results of these simulations suggest that we need a substantial reconceptualization of the pipeline for ERP analysis. It may not be productive to limit our analysis to simple summary measures like the ones studied here. Instead, the solution may be to develop a new analytic approach.

The current processing steps involve a carefully ordered process of filtering, averaging over trials, taking a summary metric, and then conducting statistical tests. While on the surface each of these steps makes sense, their statistical properties and performance in aggregate must also be considered. As we have discussed, the order of steps 3 and 4 can impact the distribution of values when working with the maximum (but not the average). Indeed, many of the substeps involve nonlinear operations and are thus not exchangeable. It is well-established (Estes, 1956; Molenaar, 1987) that the average waveform, or curve, may not be representative of the individual waveforms. A new analytic approach would be most useful if it avoided these issues of order of operations and had the flexibility to address the many moving parts of an ERP study beyond the amplitude.

One option for leveraging the hierarchy of sources of variability in ERP studies and exploiting the well-studied shapes of components is to use nonlinear mixed-effects models (although more general approaches could be considered as well, e.g., Hamaker, Dolan, & Molenaar, 2005). We used the nonlinear mixed-effects framework, Equation (4), to generate the simulations; it could be used as an analytic model as well. The framework has several advantages. The random effects at the trial level would be capable of addressing sources of trial-to-trial variability, such as background noise which may synchronize to create artificial peaks in the waveform (Yeung, Bogacz, Holroyd, & Cohen, 2004). Additionally, we would no longer need to be concerned with the order of operations during analysis because the entire process can be conducted simultaneously. It is not necessary to first average over trials—we can instead specify the hierarchy of the data (e.g., trials are nested within conditions, which are nested within individuals; or trials are nested within individuals, which are nested within conditions; or other relations across multiple channels) and the model-fitting procedure can accommodate an appropriate weighting and error structure to address temporal nonindependence. In this way, the framework can address and model violations of the IID assumption, such as temporal correlations in the ordered measurements of the EEG signal. Issues of unbalanced designs (whether from the experimental design, as a result of artifact rejection, or a result of focus on particular trials such as trials with an error response) would no longer be a concern.

More relevant to this article, the nonlinear mixed-effects model framework provides a mechanism to circumvent choosing among a set of summary metrics such as the maximum, average, or hybrid. Instead, we could make use of the functional shape of a component from which we can derive different quantifications of the waveform while controlling for other sources of variability. For example, to compare latencies for a given component across groups or conditions, we can fit Equation (4) as a nonlinear mixed-effects model, constraining L to be in a range near the anticipated time of the component (such as between 200 and 400 ms post stimulus for a P300). L is then treated as a fixed-effects term, and the remaining parameters and their uncertainty are controlled for when testing L. Further, this approach can directly model the joint distribution of amplitude and latency.

While the application discussed in this article has been ERP waveforms, the general point extends to waveform data from a wide variety of sensors where researchers are interested in assessing amplitude. We need a more careful assessment of how simple summary measures such as average amplitude or maximum amplitude in waveform data behave in the context of the complete statistical model, including the relevant preprocessing steps to prepare data for statistical testing. We found that Type I and Type II error rates varied by summary measure, various properties of the experimental design (e.g., unequal number of trials across conditions) and various properties of the data (such as variability in latency). Given that the performance of the summary measures varies by these important design and analytic details, complete and transparent reporting is important.

As demonstrated in this article a joint effort involving theoretical analysis and simulation is needed in order to gain a deeper understanding of the issues, especially given the many criteria relevant to evaluating analytic strategies such as bias, Type I error rate and Type II error rate. The continued assessment of existing practices and the development of new analytic models will become increasingly important as researchers include more sensors in their studies and collect waveform data from multiple channels.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by the University of Michigan's LS&A/Rackham/UMOR grant and the University of Michigan Institute for Social Research's GSRA Tuition Support Program.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Pamela Davis-Kean, Brian Kraus, Shuheng Zhou, the editor, and the anonymous reviewers for their comments on prior versions of this manuscript, as well as Henri Begleiter at the Neurodynamics Laboratory at the State University of New York Health Center at Brooklyn for making his data publicly available for use. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution is not intended and should not be inferred.

ORCID

Karen Nielsen https://orcid.org/0000-0003-3771-5272
Richard Gonzalez http://orcid.org/0000-0001-6334-0430

References

Arnold, B. C. (1992). Multivariate logistic distributions. In N. Balakarishnan (Ed.), *Handbook of the logistic distribution* (pp. 237–262). New York: Marcel Dekker.

Balakrishnan, N. (2013). *Handbook of the logistic distribution*. New York: Taylor & Francis.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi:10.1038/nrn3475

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, *13*(6), 407–420. doi:10.1038/nrn3241

Chew, V. (1968). Some useful alternatives to the normal distribution. *The American Statistician*, *22*(3), 22–24. doi: 10.1080/00031305.1968.10480473

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology*, *50*(2), 174–186. doi:10.1111/psyp.12001

Clayson, P. E., & Miller, G. A. (2017). ERP reliability analysis (ERA) toolbox: An open-source toolbox for analyzing the reliability of event-related brain potentials. *International Journal of Psychophysiology*, *111*, 68–79. doi: 10.1016/j.ijpsycho.2016.10.012

Cohen, M. X. (2014). *Analyzing neural time series data*. Cambridge, MA: MIT Press.

Cook, E. W. I., & Miller, G. A. (1992). Digital filtering: Background and tutorial for psychophysiologists. *Psychophysiology*, *29*(3), 350–367. doi:10.1111/j.1469-8986.1992.tb01709.x

David, H. A., & Nagaraja, H. N. (1970). *Order statistics*. Wiley Online Library.

Dien, J. (1998). Issues in the application of the average reference: Review, critiques, and recommendations. *Behavior Research Methods, Instruments, & Computers*, *30*(1), 34–43. doi:10.3758/BF03209414

Dien, J., & Santuzzi, A. (2004). Application of repeated measures ANOVA to high-density ERP datasets: A review and tutorial. In T. Handy (Ed.), *Event related potentials: A methods handbook* (pp. 57–82). Cambridge, MA: MIT Press.

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

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., … Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*, *120*(11), 1883–1908. doi:10.1016/j.clinph.2009.07.045

Esseen, C. G. (1956). A moment inequality with an application to the central limit theorem. *Scandinavian Actuarial Journal*, *1956*(2), 160–170. doi:10.1080/03461238.1956.10414946

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140. doi:10.1037/h0045156

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*(2), 180. doi:10.1017/S0305004100015681

Friedman, D., Cycowicz, Y. M., & Gaeta, H. (2001). The novelty P3: An event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neuroscience & Biobehavioral Reviews*, *25*(4), 355–373. doi:10.1016/S0149-7634(01)00019-7

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *The Annals of Mathematics*, *44*(3), 423–453. doi:10.2307/1968974

Gumbel, E. (1958). *Statistics of Extremes*. New York: Columbia University Press.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. (2005). Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis. *Multivariate Behavioral Research*, *40*(2), 207–233. doi:10.1207/s15327906mbr4002_3

Helwig, N. E. (2015). eegkit: Toolkit for electroencephalography data [Computer software manual]. *R package version 1.0-2*. Retrieved from http://CRAN.R-project.org/package=eegkit

Jennings, J. R. (1987). Editorial policy on analyses of variance with repeated measures. *Psychophysiology*, *24*(4), 474–475. doi:10.1111/j.1469-8986.1987.tb00320.x

Joyce, C., & Rossion, B. (2005). The face-sensitive N170 and VPP components manifest the same brain processes: The effect of reference electrode site. *Clinical Neurophysiology*, *116*(11), 2613–2631. doi:10.1016/j.clinph.2005.07.005

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., … Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, *51*(1), 1–21. doi:10.1111/psyp.12147

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. doi:10.1126/science.7350657

Lopez-Calderon, J., & Luck, S. J. (2014). Erplab: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213. doi:10.3389/fnhum.2014.00213

Luck, S. J. (2005). Ten simple rules for designing ERP experiments. In T. Handy (Ed.), *Event-related potentials: A methods handbook* (pp. 17–32). Cambridge, MA: The MIT Press.

Luck, S. J. (2010). Is it legitimate to compare conditions with different numbers of trials? Retrieved from http://erpinfo.org/Members/sjluck/Mean{_}Peak{_}Noise.pdf

Luck, S. J. (2014). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Makeig, S., Bell, A. J., Jung, T.-P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in Neural Information Processing Systems* (pp. 145–151). Cambridge, MA: MIT Press.

McGillem, C. D., & Aunon, J. I. (1987). Methods and analysis of brain electrical and magnetic signals. In A. Gevins & A. Rémond (Eds.), *EEG handbook* (Revised series, Vol. 1, pp. 131–170). Amsterdam: Elsevier.

Molenaar, P. (1987). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. doi:10.1207/s15366359mea0204_1

Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic ERP analyses: A step-by-step tutorial review. *Brain Topography*, *20*(4), 249–264. doi:10.1007/s10548-008-0054-5

Nielsen, K., & Gonzalez, R. (2019, July). Comparison of ERP amplitude metrics: Code. *Zenodo*. doi:10.5281/zenodo.3354464

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., … Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*(2), 127–152. doi:10.1111/1469-8986.3720127

Polich, J. (1986). P300 development from auditory stimuli. *Psychophysiology*, *23*(5), 590–597.

Polich, J., & Kok, A. (1995). Cognitive and biological determinants of P300: An integrative review. *Biological Psychology*, *41*(2), 103–146. doi:10.1016/0301-0511(95)05130-9

Proakis, J., & Manolakis, D. (1992). *Digital signal processing: Principles, algorithms, and applications*. New York: Macmillan Publishing Company.

Shevtsova, I. (2011). *On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands* (arXiv preprint arXiv:1111.6554(1)).

Thomas, D. G., Grice, J. W., Najm-Briscoe, R. G., & Williams Miller, J. (2010). The influence of unequal numbers of trials on comparisons of average event-related potentials. *Developmental Neuropsychology*, *26*(3), 753–774. doi:10.1207/s15326942dn2603

Umbricht, D., & Krljes, S. (2005). Mismatch negativity in schizophrenia: A meta-analysis. *Schizophrenia Research*, *76*(1), 1–23. doi:10.1016/j.schres.2004.12.002

Yeung, N., Bogacz, R., Holroyd, C. B., & Cohen, J. D. (2004). Detection of synchronized oscillations in the electroencephalogram: An evaluation of methods. *Psychophysiology*, *41*(6), 822–832. doi:10.1111/j.1469-8986.2004.00239.x

Zhang, X. L., Begleiter, H., Porjesz, B., & Litke, A. (1997). Electrophysiological evidence of memory impairment in alcoholic patients. *Biological Psychiatry*, *42*(12), 1157–1171. doi:10.1016/S0006-3223(96)00552-5