Received: 2 August 2022 | Revised: 4 November 2022 | Accepted: 11 January 2023

DOI: 10.1111/psyp.14363

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

Estimating statistical power for ERP studies using the auditory N1, Tb, and P2 components

Lachlan Hall | **Amy Dawel** | **Lisa-Marie Greenwood** | **Conal Monaghan** | **Kevin Berryman** | **Bradley N. Jack**

Research School of Psychology, Australian National University, Canberra, Australia

Correspondence
Bradley N. Jack, Research School of Psychology, Australian National University, Canberra, Australia.
Email: bradley.jack@anu.edu.au

Funding information
Australian Research Council, Grant/Award Number: DE220100739

Abstract

The N1, Tb, and P2 components of the event-related potential (ERP) are thought to reflect the sequential processing of auditory stimuli in the human brain. Despite their extensive use in biological, cognitive, and clinical neuroscience, there are no guidelines for how to appropriately power ERP studies using these components. In the present study, we investigated how the number of trials, number of participants, effect magnitude, and study design influenced statistical power. Using Monte Carlo simulations of ERP data from a passive listening task, we determined the probability of finding a statistically significant effect in 58,900 experiments repeated 1,000 times each. We found that as the number of trials, number of participants, and effect magnitude increased, so did statistical power. We also found that increasing the number of trials had a bigger effect on statistical power for within-subject designs than for between-subject designs, and that within-subject designs required a smaller number of trials and participants to provide the same level of statistical power for a given effect magnitude than between-subject designs. These results show that it is important to carefully consider these factors when designing ERP studies, rather than relying on tradition or anecdotal evidence. To improve the robustness and reproducibility of ERP research, we have built an online statistical power calculator (https://bradleynjack.shinyapps.io/ErpPowerCalculator), which we hope will allow researchers to estimate the statistical power of previous studies, as well as help them design appropriately-powered studies in the future.

KEYWORDS

event-related potentials (ERPs), Monte Carlo simulation, N1, P2, statistical power, Tb

1 | INTRODUCTION

The reproducibility of research findings is crucial to scientific progress. One of the most important aspects of this is statistical power, which is defined as the probability that a statistical test correctly rejects the null hypothesis when it is false, or $1-\beta$, and is usually set at 0.8 (Cohen, 1988). Statistical power is important because low-powered studies

yield more false positives (Type-I error) and false negatives (Type-II error) than high-powered studies (Button et al., 2013; Ioannidis, 2005; Ioannidis et al., 2011; Stern & Davey Smith, 2001). That is, statistical power determines the level of confidence (or lack thereof) one can have in the results of a study. Unfortunately, there are growing concerns that most of the scientific literature might be false due to a lack of statistical power (Ioannidis, 2005), and that this issue has contaminated the neuroscientific literature as well (Button et al., 2013; Smaldino & McElreath, 2016). Consistent with this concern, Clayson et al. (2019) estimated that only 15% of studies using event-related potentials (ERPs) were appropriately powered, and Larson and Carbine (2017) estimated that the majority of ERP studies do not report statistical power calculations or the information needed for others to calculate it themselves. These findings question the reproducibility of ERP research. To rectify this issue, researchers have begun to investigate how factors such as the number of trials, number of participants, effect magnitude, and study design influences the statistical power of ERP studies (Boudewyn et al., 2018; Gibney et al., 2020; Jensen & MacDonald, 2023; Ngiam et al., 2021). The aim of the present study was to investigate how these factors influence the statistical power of ERP studies using the N1, Tb, and P2 components.

The question of how many trials and participants are needed for an appropriately powered ERP study has been a topic of discussion for decades (Keil et al., 2014; Picton et al., 2000; Pivik et al., 1993). Most of this discussion has focussed on deciding the number of participants; however, there are no formal guidelines for deciding the number of trials, and it is likely that most researchers rely on tradition or anecdotal evidence (Luck, 2014). One data-driven approach has been to randomly sample a subset of trials from a larger dataset and then to quantify the similarities between the components of the resulting waveforms. This process is then repeated many times to determine the minimum number of trials required to obtain a component that is as reliable as the component from the larger dataset (Cohen & Polich, 1997; Duncan et al., 2009; Fischer et al., 2017; Huffmeijer et al., 2014; Larson et al., 2010; Marco-Pallares et al., 2011; Olvet & Hajcak, 2009; Pontifex et al., 2010; Rietdijk et al., 2014; Segalowitz & Barnes, 1993; Steele et al., 2016; Thigpen et al., 2017). However, as noted by Gehring et al. (2012) and others, the goal of most ERP studies is not to determine whether a component is present or not; rather, it is to determine whether a component differs between experimental conditions and/or groups. Furthermore, the range in which a component can be modulated by an independent variable is usually much smaller than the component itself (Luck, 2014), meaning that this approach cannot be used to determine statistical power. Therefore, it is crucial that researchers investigate how experimental factors influence statistical power.

Impact Statement

Using Monte Carlo simulations of ERP data from a passive listening task, we found that the number of trials, number of participants, effect magnitude, and study design interacted to influence statistical power for the N1, Tb, and P2 components. We hope that these results will improve the robustness and reproducibility of ERP research.

This issue was recently addressed by Boudewyn et al. (2018). In their study, they systematically manipulated the number of trials, number of participants, effect magnitude, and study design with Monte Carlo simulations of ERP data from an Eriksen flanker task, which they then repeated 1,000 times each. Using this approach, they were able to determine statistical power for the lateralized readiness potential (LRP), which is associated with the selection and preparation of a lateralized motor response (Eimer, 1998; Smulders & Miller, 2012; Vaughan et al., 1968), and the error-related negativity (ERN), which is associated with an incorrect motor response (Falkenstein et al., 1990; Gehring et al., 1993; Gehring et al., 2012). A similar approach was used by Gibney et al. (2020), who used a picture-viewing task to investigate the late positive potential (LPP), which is associated with emotional processing (Cacioppo et al., 1993; Hajcak et al., 2012; Schupp et al., 2000), and by Ngiam et al. (2021), who used a change-detection task to investigate the contralateral delay activity (CDA), which is associated with visual working memory (Luria et al., 2016; Perez & Vogel, 2012; Vogel & Machizawa, 2004). This approach was also used by Jensen and MacDonald (2023), who used the ERP CORE resource (Kappenman et al., 2021) to investigate the LRP, ERN, and five other widely used components: the N170, which is associated with face perception (Bentin et al., 1996; Bötzel & Grüsser, 1989; Rossion & Jacques, 2012); the mismatch negativity (MMN), which is associated with oddball detection (Garrido et al., 2009; Näätänen et al., 1978; Näätänen & Kreegipuu, 1987); the N2pc, which is associated with selective attention (Luck, 2012; Luck & Hillyard, 1990, 1994); the N400, which is associated with semantic processing (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980; Swaab et al., 2012); and the P3, which is associated with decision-making (Chapman & Bragdon, 1964; Polich, 2007, 2012). Separately, these studies found that as the number of trials, number of participants, and effect magnitude increased, so did statistical power. They also found that increasing the number of trials had a bigger effect on statistical power for within-subject designs than for between-subject

designs, and that within-subject designs required a smaller number of trials and participants to provide the same level of statistical power for a given effect magnitude than between-subject designs. Even though these results are informative, these studies provided different recommendations for each of the different components. This is important because it suggests that specific recommendations for one component are unlikely to generalize to other components.

In the present study, we sought to determine statistical power for the N1, Tb, and P2 components of the ERP. These components are elicited by an auditory stimulus, they are observed one after the other in the ERP waveform, and they have their neural sources in increasingly higher cortical regions (Crowley & Colrain, 2004; Joos et al., 2014; Näätänen & Picton, 1987; Woods, 1995). As such, they are thought to reflect the sequential processing of auditory stimuli in the human brain (Picton, 2010). We chose these components because, unlike those used by Boudewyn et al. (2018), Gibney et al. (2020), Jensen and MacDonald (2023), and Ngiam et al. (2021), the N1, Tb, and P2 are typically smaller in amplitude, earlier in latency after stimulus onset, and are assumed to require more trials to isolate; thus, their characteristics are very different from those used in previous research. We also chose these components because they are widely used in biological and cognitive neuroscience to study perception (Mulert et al., 2005), attention (Näätänen, 1992), learning (Tremblay et al., 2014), expertise (Shahin et al., 2003), and language (Steinhauer & Connolly, 2008), as well as in clinical neuroscience to study auditory development (Bishop et al., 2011), hearing thresholds (Campbell & Muller-Gass, 2011), tinnitus (Lee et al., 2007), and schizophrenia (Salisbury et al., 2010). Despite their extensive use, there are no guidelines for how to appropriately power ERP studies using these components. To address this, we used Monte Carlo simulations of ERP data from a passive listening task to investigate how the number of trials, number of participants, effect magnitude, and study design influences the statistical power of the N1, Tb, and P2 components.

2 | METHOD

2.1 | Participants

Fifty students from the Australian National University (ANU) participated in our experiment for course credit. All participants gave written informed consent prior to the experiment and reported having normal hearing in both ears. Mean age of the participants, 29 of whom were female and 45 of whom were right-handed, was 21 ($SD = 2$) years. The experiment was approved by the ANU Science and Medical Delegated Ethics Review Committee and was conducted in accordance with the ethical standards laid down in the Declaration of Helsinki (World Medical Association, 2013).

2.2 | Apparatus, stimuli, and procedure

Participants completed a passive listening task: they were instructed to ignore 1,000 identical sinusoidal tones presented binaurally through headphones (Audio-Technica ATH-M20x) while watching a silent film on a computer monitor (Dell U2415). The tones had a frequency of 1,000 Hz, a duration of 100 ms including 5-ms rise and fall times, and an intensity of 75 dB SPL. The stimulus-onset asynchrony between any two tones randomly varied between 1,000 and 2,000 ms (rectangular distribution, average of 1,500 ms). The films consisted of a selection of nature documentaries, were silent (the sound was muted), and did not contain subtitles. Stimulus presentation was controlled by specially written Matlab scripts using the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). The task took about 25 minutes to complete.

2.3 | Electroencephalogram (EEG) acquisition

We recorded the EEG with a BioSemi ActiveTwo system using 64 Ag/AgCl active electrodes placed according to the extended 10–20 system (FP1, FPz, FP2, AF7, AF3, AFz, AF4, AF8, F7, F5, F3, F1, Fz, F2, F4, F6, F8, FT7, FC5, FC3, FC1, FCz, FC2, FC4, FC6, FT8, T7, C5, C3, C1, Cz, C2, C4, C6, T8, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, P9, P7, P5, P3, P1, Pz, P2, P4, P6, P8, P10, PO7, PO3, POz, PO4, PO8, O1, Oz, O2, and Iz). We recorded the vertical electrooculogram (EOG) by placing an electrode above (we used FP1) and below the left eye and the horizontal EOG by placing an electrode on the outer canthus of each eye. We also placed an electrode on the tip of the nose. The EEG was sampled at 1,024 Hz.

2.4 | EEG processing and region of interest (ROI) selection

We re-referenced the EEG data to the electrode on the tip of the nose, and we filtered the data using a half-amplitude 0.1–30 Hz phase-shift free Butterworth filter (12 dB/Oct slope), as well as a 50-Hz Notch filter. We extracted the epochs from −100 ms to 400 ms relative to sound onset, we corrected the epochs for eye-blink and movement artifacts

using the technique described in Gratton et al. (1983) and Miller et al. (1988), and we excluded all epochs with signals exceeding peak-to-peak amplitudes of 200 μV at any EEG channel. We baseline-corrected all epochs to their mean voltage from −100 to 0 ms, and we computed an ERP waveform for each participant from the remaining trials. On average, the waveform was computed from 965 ($SD = 27$) artifact-free epochs. We computed a grand-average ERP waveform, and we analyzed the N1 at fronto-central (Fz, FCz, and Cz) electrodes in the time-window of 84–124 ms, the Tb at bilateral temporal (T7 and T8) electrodes in the time-window of 124–164 ms, and the P2 at central (FCz, Cz, and CPz) electrodes in the time-window of 151–191 ms. We chose these electrodes to be consistent with those in the literature (Crowley & Colrain, 2004; Näätänen & Picton, 1987; Woods, 1995), and we chose these time-windows by centering a 40-ms time-window around each peak on the grand-averaged ERP waveform (Luck & Gaspelin, 2017).

2.5 | Quantifying the noise

We used three techniques for quantifying the noise in our data. First, we computed amplitude density using the Fast Fourier Transform (FFT). Specifically, after re-referencing and filtering the data, we segmented the data into 5-s epochs with 50% overlap, we excluded all epochs with signals exceeding peak-to-peak amplitudes of 200 μV at any EEG channel, and we computed the amplitude density at each frequency from 1–100 Hz in 0.125 Hz steps using the FFT. The amplitude density was averaged across epochs, electrodes, and participants. Figure 1 shows the grand-averaged amplitude density spectrum. Second, we used the plus–minus averaging technique described in Schimmel (1967). This technique removes the ERP signal while leaving the noise by subtracting the ERP for odd-numbered trials from the ERP for even-numbered trials for each participant. Third, we used the standardized measurement error (SME) technique described in Luck et al. (2020). This technique is similar to the standard error of measurement, except that it can be applied to a specific time-window for each participant, yielding a measure of precision, and then aggregated across participants to provide a measure of the quality of the data.

2.6 | Monte Carlo simulation and statistical analysis

We used Monte Carlo methods to simulate a large number of separate experiments with different parameters by randomly sampling (with replacement) a subset of trials and participants from the dataset described above. Each experiment included a specific number of trials, which ranged from 20–1,000 in increments of 20 trials; number of participants, which ranged from 10–100 in increments of five participants; and effect magnitude, which ranged from 0–3 in increments of 0.1 μV, and used either a within- or between-subject design. Combining these parameters led to a total of 58,900 experiments, each of which was repeated 1000 times. For the experiments using a within-subject design, we sampled twice the number of trials from each participant to simulate two experimental conditions, we added half of the effect magnitude to one condition and subtracted half of the effect magnitude from the other condition, and we tested for a statistically significant ($\alpha = .05$) effect between the two conditions at the N1, Tb, and P2 using separate paired-samples *t*-tests (two-tailed). For the experiments using a between-subject design, we sampled twice the number of participants to simulate two experimental groups, we added half of the effect magnitude to one group and subtracted half of the effect magnitude from the other group, and we tested for a statistically significant effect between the two groups at the N1, Tb, and P2 using separate independent-samples *t*-tests (two-tailed). This approach is ideal for our purpose because it combines real ERP data with artificially induced experimental effects. To calculate statistical power, we divided the number of significant experiments for a given set of parameters by the total number of repetitions (i.e., 1,000).

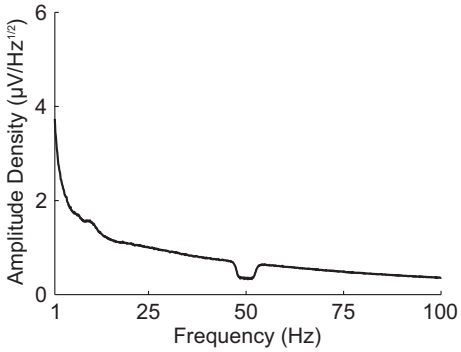

FIGURE 1 Amplitude density as a function of frequency, calculated from FFTs of data averaged across epochs, electrodes, and participants.

3 | RESULTS

3.1 | ERP and noise waveforms

Figure 2a shows the grand-averaged ERP waveform as well as the individual and grand-averaged plus-minus waveforms at fronto-central electrodes. Consistent with the N1 literature (Näätänen & Picton, 1987;

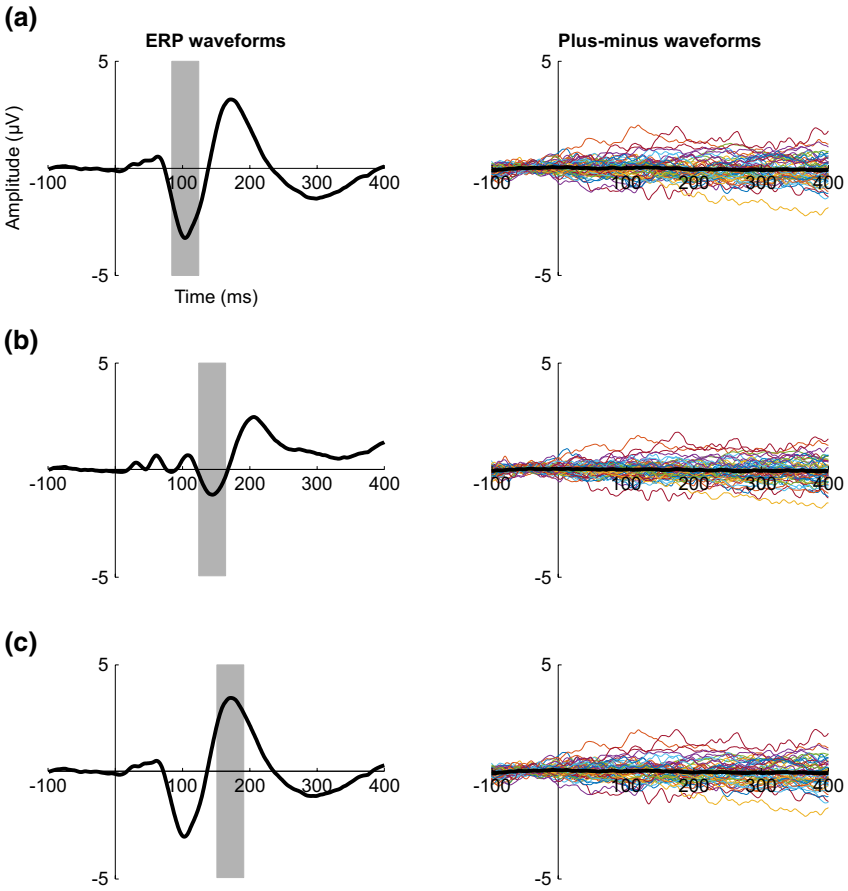

FIGURE 2 ERP waveforms and noise. The left panels show the grand-averaged ERP waveforms at (a) fronto-central (Fz, FCz, and Cz), (b) bilateral temporal (T7 and T8), and (c) central electrodes (FCz, Cz, and CPz), showing time (ms) on the *x*-axis, with 0 indicating sound onset, and voltage (μV) on the *y*-axis, with positive voltages plotted upwards. The gray bars show the (a) N1, (b) Tb, and (c) P2 time-windows. The right panels show the individual and grand-averaged plus-minus waveforms at (a) fronto-central, (b) bilateral temporal, and (c) central electrodes.

Woods, 1995), the grand-averaged ERP waveform shows a negative-going deflection starting at about 75 ms, peaking at 104 ms, and returning to baseline at about 135 ms. Figure 2b shows the waveforms at bilateral temporal electrodes. Consistent with the Tb literature (Näätänen & Picton, 1987; Woods, 1995), the grand-averaged ERP waveform shows a negative-going deflection starting at about 120 ms, peaking at 144 ms, and returning to baseline at about 160 ms. Figure 2c shows the waveforms at central electrodes. Consistent with the P2 literature (Crowley & Colrain, 2004), the grand-averaged ERP waveform shows a negative-going deflection starting at about 135 ms, peaking at 171 ms, and returning to baseline at about 230 ms. The grand-averaged plus-minus waveforms at fronto-central, bilateral temporal, and central electrodes were approximately 0 for the duration of the epoch, indicating that the ERP signals were successfully removed by the plus-–minus averaging technique, leaving only the noise (Schimmel, 1967). The SMEs for the N1, Tb, and P2 time-windows were 0.59, 0.55, and 0.71, respectively, which were much smaller than the *SD*s, 1.64, 1.14, and 2.01, respectively, indicating that the contribution of measurement error to the observed variability across participants was not as great as the contribution of true differences among participants (Luck et al., 2020). That is, the differences across individual participants is driven by true individual differences, rather than by measurement error or poor quality of data.

3.2 | N1 simulations

Figure 3a shows the probability of obtaining a statistically significant N1 effect for within-subject designs. Consistent with previous research (Boudewyn et al., 2018; Gibney et al., 2020; Jensen & MacDonald, 2023; Ngiam et al., 2021), we found that the number of trials, number of participants, effect magnitude, and study design interacted to influence statistical power. For instance, for an effect magnitude of 0.5 μV, if there were 300 trials, then 75 participants were needed to obtain appropriate statistical power (which we defined as 0.8, as is the norm in neuroscience; Button et al., 2013; Smaldino & McElreath, 2016), and if the number of trials was doubled to 600, then the number of participants was reduced to 40. For an effect magnitude of 1 μV, if there were 200 trials, then 30 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 400, then the number of participants was reduced to 15; and if the number of trials was doubled again to 800, then the number of participants was reduced to 10. For an effect magnitude of 1.5 μV, if there were 100 trials, then 25 participants

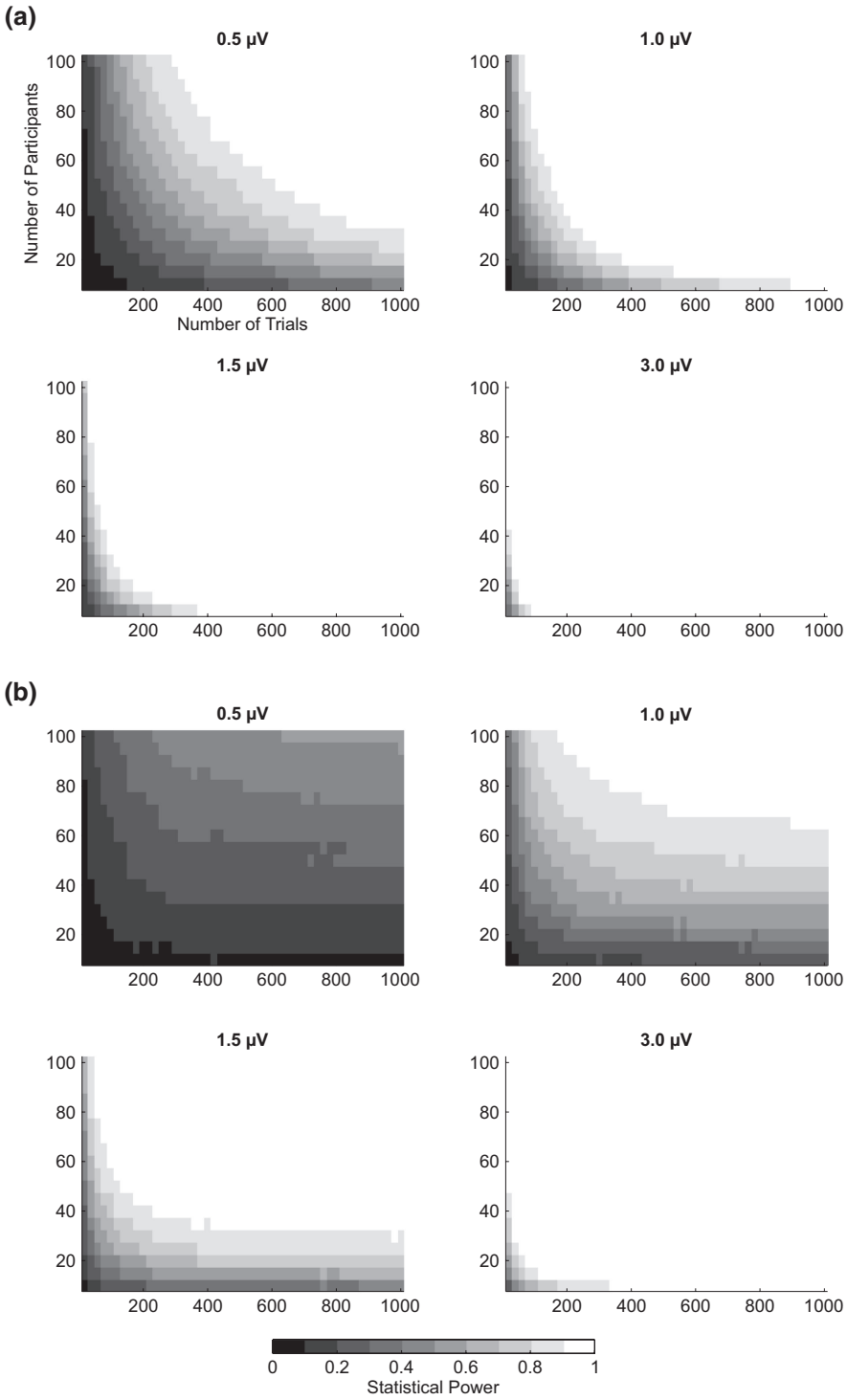

FIGURE 3 N1 simulations. The distribution plots show the probability of obtaining a statistically significant ($\alpha = .05$) N1 effect as a function of the number of trials and participants for effect magnitudes of 0.5, 1, 1.5, and 3 μV, which are typical of those reported in the N1 literature, for (a) within- and (b) between-subject designs.

were needed to obtain appropriate statistical power; if the number of trials was doubled to 200, then the number of participants was reduced to 15; and if the number of trials was doubled again to 400, then the number of participants was reduced to 10. For an effect magnitude of 3 μV, if there were 40 trials, then 20 participants were needed to obtain appropriate statistical power, and if the number of trials was doubled to 80, then the number of participants was reduced to 10.

Figure 3b shows the probability of obtaining a statistically significant N1 effect for between-subject designs. For an effect magnitude of 0.5 μV, even 1,000 trials and 100 participants were insufficient to obtain appropriate statistical power. For an effect magnitude of 1 μV, if there were 200 trials, then 70 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 400, then the number of participants was reduced to 60; and if the number of trials was

doubled again to 800, then the number of participants was reduced to 50. For an effect magnitude of 1.5 μV, if there were 100 trials, then 50 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 200, then the number of participants was reduced to 35; and if the number of trials was doubled again to 400, then the number of participants was reduced to 25. For an effect magnitude of 3 μV, if there were 40 trials, then 25 participants were needed to obtain appropriate statistical power, and if the number of trials was doubled to 80, then the number of participants was reduced to 15.

3.3 | Tb simulations

Figure 4a shows the probability of obtaining a statistically significant Tb effect for within-subject designs. For an effect magnitude of 0.5 μV, if there were 300 trials, then 65 participants were needed to obtain appropriate statistical power, and if the number of trials was doubled to 600, then the number of participants was reduced to 35. For an effect magnitude of 1 μV, if there were 200 trials, then 25 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 400, then the number of participants was reduced to 15; and if the number of trials was doubled again to 800, then the number of participants was reduced to 10. For an effect magnitude of 1.5 μV, if there were 100 trials, then 25 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 200, then the number of participants was reduced to 15; and if the number of trials was doubled again to 400, then the number of participants was reduced to 10. For an effect magnitude of 3 μV, if there were 40 trials, then 15 participants were needed to obtain appropriate statistical power, and if the number of trials was doubled to 80, then the number of participants was reduced to 10.

Figure 4b shows the probability of obtaining a statistically significant Tb effect for between-subject designs. For an effect magnitude of 0.5 μV, 880 trials and 100 participants were needed to obtain appropriate statistical power, but increasing the number of trials to 1,000 did not reduce the number of participants. For an effect magnitude of 1 μV, if there were 200 trials, then 45 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 400, then the number of participants was reduced to 35; and if the number of trials was doubled again to 800, then the number of participants was reduced to 30. For an effect magnitude of 1.5 μV, if there were 100 trials, then 30 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 200, then the number of participants was reduced to 20; and if the number of trials was doubled again to 400, then the number of participants was reduced to 15. For an effect magnitude of 3 μV, if there were 40 trials, then 20 participants were needed to obtain appropriate statistical power, and if the number of trials was doubled to 80, then the number of participants was reduced to 15.

3.4 | P2 simulations

Figure 5a shows the probability of obtaining a statistically significant P2 effect for within-subject designs. For an effect magnitude of 0.5 μV, if there were 300 trials, then more than 100 participants were needed to obtain appropriate statistical power, and if the number of trials was doubled to 600, then the number of participants was reduced to 55. For an effect magnitude of 1 μV, if there were 200 trials, then 45 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 400, then the number of participants was reduced to 20; and if the number of trials was doubled again to 800, then the number of participants was reduced to 10. For an effect magnitude of 1.5 μV, if there were 100 trials, then 40 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 200, then the number of participants was reduced to 20; and if the number of trials was doubled again to 400, then the number of participants was reduced to 15. For an effect magnitude of 3 μV, if there were 40 trials, then 25 participants were needed to obtain appropriate statistical power, and if the number of trials was doubled to 80, then the number of participants was reduced to 15.

Figure 5b shows the probability of obtaining a statistically significant P2 effect for between-subject designs. For an effect magnitude of 0.5 μV, even 1,000 trials and 100 participants were insufficient to obtain appropriate statistical power. For an effect magnitude of 1 μV, if there were 200 trials, then 100 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 400, then the number of participants was reduced to 80; and if the number of trials was doubled again to 800, then the number of participants was reduced to 75. For an effect magnitude of 1.5 μV, if there were 100 trials, then 65 participants were needed to obtain appropriate statistical power; if the number of trials was doubled to 200, then the number of participants was reduced to 50; and if the number of trials was doubled again to 400, then the number of participants was reduced to 40. For an effect magnitude of 3 μV, if there were 40 trials, then 30 participants were needed to obtain appropriate statistical power, and if the number of trials was doubled to 80, then the number of participants was reduced to 20.

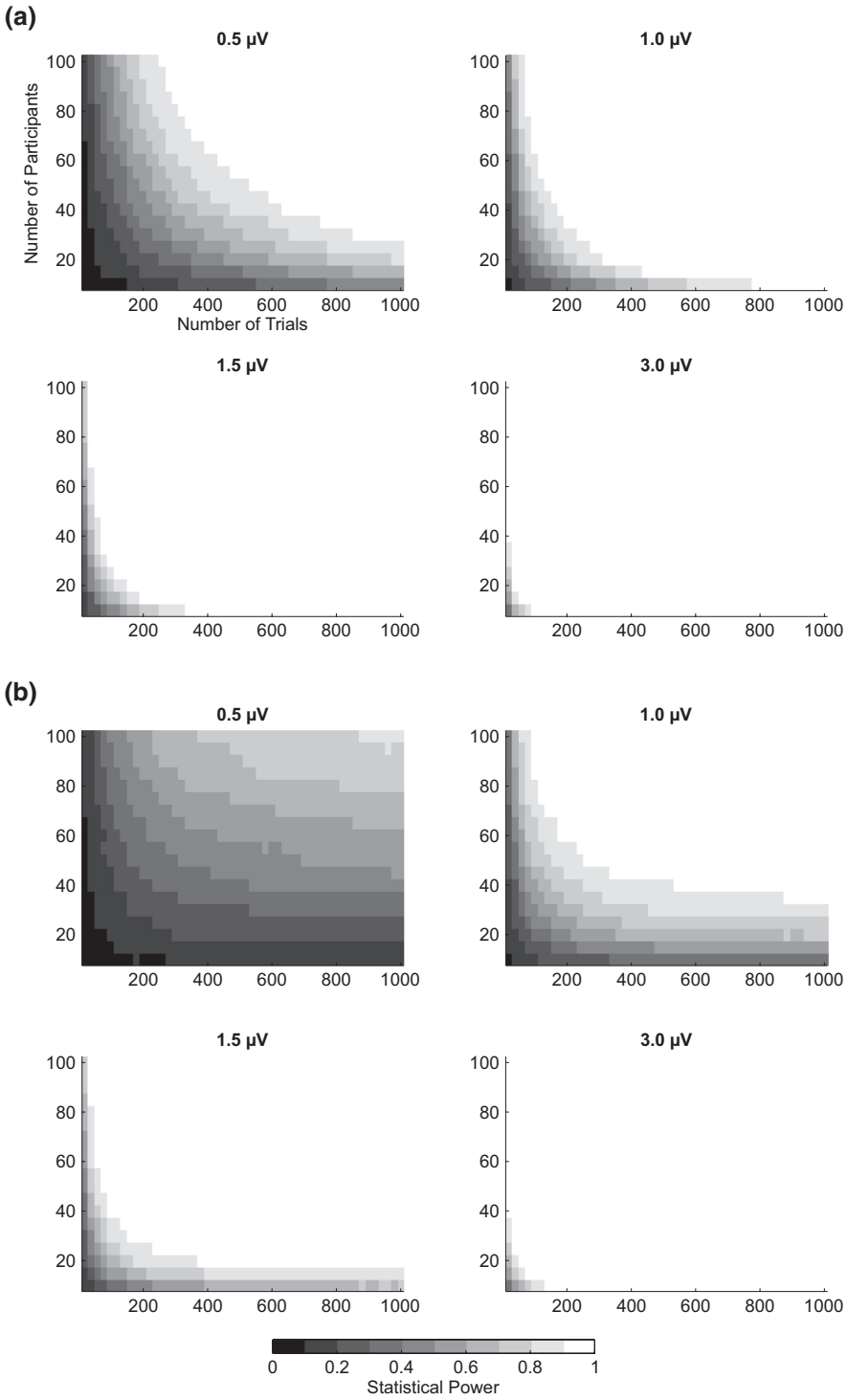

FIGURE 4 Tb simulations. The distribution plots show the probability of obtaining a statistically significant ($\alpha = .05$) Tb effect as a function of the number of trials and participants for effect magnitudes of 0.5, 1, 1.5, and 3 μV, which are typical of those reported in the Tb literature, for (a) within- and (b) between-subject designs.

4 | DISCUSSION

In the present study, we sought to determine statistical power for the N1, Tb, and P2 components of the ERP. To accomplish this, we used Monte Carlo simulations of ERP data from a passive listening task to systematically manipulate the number of trials, number of participants, effect magnitude, and study design, resulting in 58,900 experiments which we then repeated 1,000 times each. Consistent with Boudewyn et al. (2018), Gibney et al. (2020), Jensen and MacDonald (2023), and Ngiam et al. (2021), we found that the number of trials, number of participants, effect magnitude, and study design interacted to influence statistical power in at least three ways. First, we found that as the number of trials, number of participants, and effect magnitude increased, so did statistical power. This can be seen in Figure 3a, where statistical power for a given number

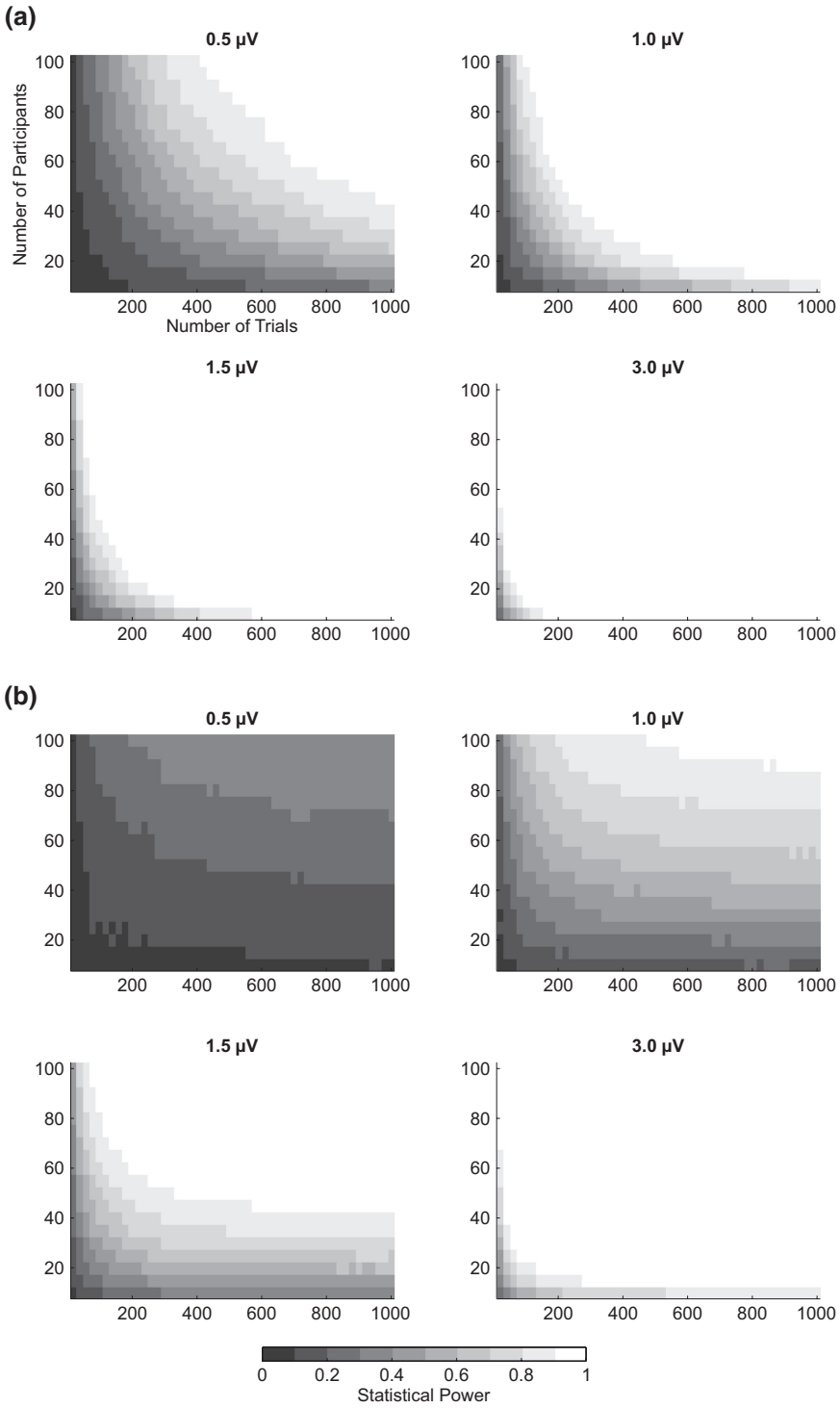

FIGURE 5 P2 simulations. The distribution plots show the probability of obtaining a statistically significant ($\alpha = .05$) P2 effect as a function of the number of trials and participants for effect magnitudes of 0.5, 1, 1.5, and 3 μV, which are typical of those reported in the P2 literature, for (a) within- and (b) between-subject designs.

of trials, number of participants, and effect magnitude increased as any one of these factors also increased. This is an important message for the neuroscience community because it is common for researchers to estimate (usually by relying on tradition or anecdotal evidence; Luck, 2014) the statistical power of a study based on its number of participants, such that studies with a small number of participants are assumed to have low statistical power, whereas studies with a large number of participants are assumed to have high statistical power. However, our results show that it is possible for a study with a small number of participants to have high statistical power if it also has a large number of trials and/or a large effect magnitude and that it is possible for a study with a large number of participants to have low statistical power if it also has a small number of trials and/or a small effect magnitude. That is, our results show how studies with a small number of participants can, on

some occasions, have more statistical power than studies with a large number of participants.

Second, we found that increasing the number of trials had a bigger effect on statistical power for within-subject designs than for between-subject designs. This can be seen by comparing Figure 3a with the corresponding plots in Figure 3b, with the former tending to reach the minimum number of participants more often than the latter, especially for effect magnitudes smaller than or equal to 1.5 μV. Third, we found that within-subject designs required a smaller number of trials and participants to provide the same level of statistical power for a given effect magnitude than between-subject designs. This can be seen by comparing Figure 3a with the corresponding plots in Figure 3b, with the former tending to provide more examples of different combinations of the number of trials and participants in which statistical power is equal to or larger than 0.8 than the latter, especially for effect magnitudes smaller than or equal to 1.5 μV. Importantly, our characterization of the results is not specific to Figure 3, which shows the N1 simulations; we can see the same patterns in Figures 4 and 5, which show the Tb and P2 simulations, respectively. Similar to Boudewyn et al. (2018), we suspect that the reason for the differences between within- and between-subject designs is the main source of variance: if the main source of variance is the number of trials, as is often the case in within-subject designs, then increasing the number of trials should decrease the variance and therefore increase statistical power. If, however, the main source of variance is individual differences, as is often the case in between-subject designs, then increasing the number of trials should have a smaller effect on decreasing the variance and increasing statistical power.

As mentioned above, the key finding of the present study is that the number of trials, number of participants, effect magnitude, and study design interacted to influence statistical power. Consistent with previous research (Boudewyn et al., 2018; Gibney et al., 2020; Jensen & MacDonald, 2023; Ngiam et al., 2021), this shows that there is no single answer to the question of how many trials or participants are needed for an appropriately powered ERP study. Instead, the number of trials required to obtain appropriate statistical power depends on the number of participants, effect magnitude, and study design. Similarly, the number of participants required to obtain appropriate statistical power depends on the number of trials, effect magnitude, and study design. Intriguingly, it appears as though the statistical power of an ERP study is also influenced by the component of interest. By consolidating the results of Boudewyn et al. (2018), Gibney et al. (2020), Jensen and MacDonald (2023), and Ngiam et al. (2021), we noticed that they provided different recommendations for each of the different components. A key difference between these studies, which could explain the different recommendations, is that they investigated different components with different characteristics. Consistent with this, in the present study, we found that the Tb required a marginally smaller number of trials and participants than the N1 for a given effect magnitude and study design, which required a marginally smaller number of trials and participants than the P2 for a given effect magnitude and study design. This can be seen by comparing Figures 3-5, which show the N1, Tb, and P2 simulations, respectively. This suggests that our recommendations, as well as those of Boudewyn et al. (2018), Gibney et al. (2020), Jensen and MacDonald (2023), and Ngiam et al. (2021), might not generalize to other components. Because of this, we strongly encourage researchers to adopt our data-driven approach to estimating statistical power for their component(s) of interest or use alternative methods for estimating statistical power, such as Baker et al.'s (2021) method, until data for their component(s) of interest become available.

Even though there is no single answer to the question of how many trials or participants are needed for an appropriately powered ERP study, our results might be useful for improving the robustness and reproducibility of ERP research (Garrett-Ruffin et al., 2021; Kappenman & Keil, 2017; Larson & Moser, 2017; Pavlov et al., 2021). To facilitate this, we have built an online statistical power calculator (https://bradleynjack.shinyapps.io/ErpPowerCalculator). We encourage researchers interested in the N1, Tb, and/or P2 to use this calculator to estimate the statistical power of previous studies, as well as help them design appropriately powered studies in the future. Of course, it is important to acknowledge that even though $\alpha = .05$ is the norm in neuroscience (Button et al., 2013; Smaldino & McElreath, 2016), our recommendations will not apply to experiments using a different threshold. To help researchers using a different threshold (Benjamin et al., 2018; Lakens et al., 2018; Maier & Lakens, 2022; Miller & Ulrich, 2019), our calculator also estimates statistical power when $\alpha = .01$, .005, and .001. Relatedly, our recommendations should be treated with some caution as they might not generalize to situations that are significantly different to our dataset. For example, it is unclear whether our recommendations will apply to these components when elicited by different populations, stimuli, or paradigms (Kappenman & Luck, 2017; Picton, 2010; Puce & Hämäläinen, 2017), recorded in different environments or by different EEG or electrode systems (Kappenman & Luck, 2010; Laszlo et al., 2014), or computed by different processing and analytical pipelines (Clayson et al., 2021; Sandre et al., 2020). Despite these limitations, our recommendations are more informative for

estimating statistical power than relying on tradition or anecdotal evidence, and as such, may provide a platform for the development of psychological theory (Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019) and ERP biomarkers of clinical disorders (Luck et al., 2011).

AUTHOR CONTRIBUTIONS

Lachlan Hall: Conceptualization; investigation; methodology; software; writing – original draft. **Amy Dawel:** Supervision; writing – review and editing. **Lisa-Marie Greenwood:** Supervision; writing – review and editing. **Conal Monaghan:** Software; writing – review and editing. **Kevin Berryman:** Data curation; writing – review and editing. **Bradley N. Jack:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; validation; visualization; writing – original draft; writing – review and editing.

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council (DE220100739). Open access publishing facilitated by Australian National University, as part of the Wiley - Australian National University agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on reasonable request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Bradley N. Jack https://orcid.org/0000-0003-0523-6656

REFERENCES

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*, 295–314. https://doi.org/10.1037/met0000337

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bentin, S., McCarthy, G., Perez, E., Puce, A., & Allison, T. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, *8*, 551–565. https://doi.org/10.1162/jocn.1996.8.6.551

Bishop, D. V. M., Anderson, M., Reid, C., & Fox, A. M. (2011). Auditory development between 7 and 11 years: An event-related potential (ERP) study. *PLoS One*, *6*(e18993), 1–11. https://doi.org/10.1371/journal.pone.0018993

Bötzel, K., & Grüsser, O. J. (1989). Electric brain potentials evoked by pictures of faces and non-faces: A search for face-specific EEG-potentials. *Experimental Brain Research*, *77*, 349–360. https://doi.org/10.1007/bf00274992

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, *55*(e13049), 1–16. https://doi.org/10.1111/psyp.13049

Brainard, D. H. (1997). The Psychophysics tool box. *Spatial Vision*, *10*, 433–436. https://doi.org/10.1163/156856897x00357

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

Cacioppo, J. T., Crites, S. L., Berntson, G. G., & Coles, M. G. (1993). If attitudes affect how stimuli are processed, should they not affect the event-related brain potential? *Psychological Science*, *4*, 108–112. https://doi.org/10.1111/j.1467-9280.1993.tb00470.x

Campbell, K., & Muller-Gass, A. (2011). The extent of processing of near-hearing threshold stimuli during natural sleep. *Sleep*, *34*, 1243–1249. https://doi.org/10.5665/sleep.1248

Chapman, R. M., & Bragdon, H. R. (1964). Evoked responses to numerical and non-numerical visual stimuli while problem solving. *Nature*, *203*, 1155–1157. https://doi.org/10.1038/2031155a0

Clayson, P. E., Baldwin, S., Rocha, H. A., & Larson, M. J. (2021). The data-processing multiverse of event-related potentials (ERPs): A roadmap for the optimization and standardization of ERP processing and reduction pipelines. *NeuroImage*, *245*, 118712. https://doi.org/10.1016/j.neuroimage.2021.118712

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(e13437), 1–17. https://doi.org/10.1111/psyp.13437

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic. https://doi.org/10.4324/9780203771587

Cohen, J., & Polich, J. (1997). On the number of trials needed for P300. *International Journal of Psychophysiology*, *25*, 249–255. https://doi.org/10.1016/s0167-8760(96)00743-x

Crowley, K. E., & Colrain, I. M. (2004). A review of the evidence for P2 being an independent component process: Age, sleep and modality. *Clinical Neurophysiology*, *115*, 732–744. https://doi.org/10.1016/j.clinph.2003.11.021

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

Eimer, M. (1998). The lateralized readiness potential as an on-line measure of central response activation processes. *Behavior Research Methods, Instruments, & Computers*, *30*, 145–156. https://doi.org/10.1016/s0166-4115(97)80027-1

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*, 1–10. https://doi.org/10.1177/1745691620970586

Falkenstein, M., Hohnsbein, J., & Hoormann, J. (1990). Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In C. H. M. Brunia, A. W. K. Gaillard, & A. Kok (Eds.), *Psychophysiological brain research* (Vol. *78*, pp. 447–455). Tilburg University Press. https://doi.org/10.1016/0013-4694(91)90062-9

Fischer, A. G., Klein, T. A., & Ullsperger, M. (2017). Comparing the error-related negativity across groups: The impact of error- and trial-number differences. *Psychophysiology*, *54*, 998–1009. https://doi.org/10.1111/psyp.12863

Garrett-Ruffin, S., Hindash, A. C., Kaczkurkin, A. N., Mears, R. P., Morales, S., Paul, K., Pavlov, Y. G., & Keil, A. (2021). Open science in psychophysiology: An overview of challenges and emerging solutions. *International Journal of Psychophysiology*, *162*, 69–78. https://doi.org/10.1016/j.ijpsycho.2021.02.005

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology*, *120*, 453–463. https://doi.org/10.1016/j.clinph.2008.11.029

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*, 385–390. https://doi.org/10.1111/j.1467-9280.1993.tb00586.x

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2012). The error-related negativity (ERN/Ne). In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0120

Gibney, K. D., Kypriotakis, G., Cinciripini, P. M., Robinson, J. D., Minnix, J. A., & Versace, F. (2020). Estimating statistical power for event-related potential studies using the late positive potential. *Psychophysiology*, *57*(e13482), 1–15. https://doi.org/10.1111/psyp.13482

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*, 468–484. https://doi.org/10.1016/0013-4694(83)90135-9

Hajcak, G., Weinberg, A., MacNamara, A., & Foti, D. (2012). ERPs and the study of emotion. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0222

Huffmeijer, R., Bakermans-Kranenburg, M. J., Alink, L. R., & van Ijzendoorn, M. H. (2014). Reliability of event-related potentials: The influence of number of trials and electrodes. *Physiology & Behavior*, *130*, 13–22. https://doi.org/10.1016/j.physbeh.2014.03.008

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(e124), 1–6. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A., Tarone, R., & McLaughlin, J. K. (2011). The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*, *22*, 450–456. https://doi.org/10.1097/ede.0b013e31821b506e

Jensen, K. M., & MacDonald, J. A. (2023). Towards thoughtful planning of ERP studies: How participants, trials, and effect magnitude interact to influence statistical power across seven ERP components. *Psychophysiology*, *61*(e14245), 1–26. https://doi.org/10.1111/psyp.14245

Joos, K., Gilles, A., Van de Heyning, P., De Ridder, D., & Vanneste, S. (2014). From sensation to percept: The neural signature of auditory event-related potentials. *Neuroscience & Biobehavioral Reviews*, *42*, 148–156. https://doi.org/10.1016/j.neubiorev.2014.02.009

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *NeuroImage*, *225*(117465), 1–12. https://doi.org/10.1016/j.neuroimage.2020.117465

Kappenman, E. S., & Keil, A. (2017). Introduction to the special issue on recentering science: Replication, robustness, and reproducibility in psychophysiology. *Psychophysiology*, *54*, 3–5. https://doi.org/10.1111/psyp.12787

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology*, *47*, 888–904. https://doi.org/10.1111/j.1469-8986.2010.01009.x

Kappenman, E. S., & Luck, S. J. (2017). Best practices for event-related potential research in clinical populations. *Biological Psychiatry*, *1*, 110–115. https://doi.org/10.1016/j.bpsc.2015.11.007

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

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*, 1–235. https://doi.org/10.1177/030100660703601s101

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

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

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168–171. https://doi.org/10.1038/s41562-018-0311-x

Larson, M. J., Baldwin, S. A., Good, D. A., & Fair, J. E. (2010). Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): The role of number of trials. *Psychophysiology*, *47*, 1167–1171. https://doi.org/10.1111/j.1469-8986.2010.01022.x

Larson, M. J., & Carbine, K. A. (2017). Sample size calculations in human electrophysiology (EEG and ERP) studies: A systematic review and recommendations for increased rigor. *International Journal of Psychophysiology*, *111*, 33–41. https://doi.org/10.1016/j.ijpsycho.2016.06.015

Larson, M. J., & Moser, J. S. (2017). Rigor and replication: Toward improved best practices in human electrophysiology research. *International Journal of Psychophysiology*, *111*, 1–4. https://doi.org/10.1016/j.ijpsycho.2016.12.001

Laszlo, S., Ruiz-Blondet, M., Khalifian, N., Chu, F., & Jin, Z. (2014). A direct comparison of active and passive amplification electrodes in the same amplifier system. *Journal of Neuroscience Methods*, *235*, 298–307. https://doi.org/10.1016/j.jneumeth.2014.05.012

Lee, C. Y., Jaw, F. S., Pan, S. L., Lin, M. Y., & Young, Y. H. (2007). Auditory cortical evoked potentials in tinnitus patients with normal audiological presentation. *Journal of the Formosan Medical Association*, *106*, 979–985. https://doi.org/10.1016/s0929-6646(08)60072-8

Luck, S. J. (2012). Electrophysiological correlates of the focusing of attention within complex visual scenes: N2pc and related ERP components. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0161

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

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*, 146–157. https://doi.org/10.1111/psyp.12639

Luck, S. J., & Hillyard, S. A. (1990). Electrophysiological evidence for parallel and serial processing during visual search. *Perception & Psychophysics*, *48*, 603–617. https://doi.org/10.3758/bf03211606

Luck, S. J., & Hillyard, S. A. (1994). Spatial filtering during visual search: Evidence from human electrophysiology. *Journal of Experimental Psychology: Human Perception & Performance*, *20*, 1000–1014. https://doi.org/10.1037//0096-1523.20.5.1000

Luck, S. J., Mathalon, D. H., O'Donnell, B. F., Hämäläinen, M. S., Spencer, K. S., Javitt, D. C., & Uhlhaas, P. J. (2011). A roadmap for the development and validation of ERP biomarkers in schizophrenia research. *Biological Psychiatry*, *70*, 28–34. https://doi.org/10.1016/j.biopsych.2010.09.021

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2020). Standardized measurement error: A universal measure of data quality for averaged event-related potentials. *Psychophysiology*, *58*(e13793), 1–15. https://doi.org/10.1111/psyp.13793

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews*, *62*, 100–108. https://doi.org/10.1016/j.neubiorev.2016.01.003

Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2), 1–14. https://doi.org/10.1177/25152459221080396

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, *48*, 852–860. https://doi.org/10.1111/j.1469-8986.2010.01152.x

Miller, G. A., Gratton, G., & Yee, C. M. (1988). Generalized implementation of an eye movement correction procedure. *Psychophysiology*, *25*, 241–243. https://doi.org/10.1111/j.1469-8986.1988.tb00999.x

Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLoS One*, *14*(e0208631), 1–13. https://doi.org/10.1371/journal.pone.0208631

Mulert, C., Jäger, L., Propp, S., Karch, S., Störmann, S., Pogarell, O., Möller, H.-J., Juckel, G., & Hegerl, U. (2005). Sound level dependence of the primary auditory cortex: Simultaneous measurement with 61-channel EEG and fMRI. *NeuroImage*, *28*, 49–58. https://doi.org/10.1016/j.neuroimage.2005.05.041

Näätänen, R. (1992). *Attention and brain function*. Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9780429487354

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*, 313–329. https://doi.org/10.1016/0001-6918(78)90006-9

Näätänen, R., & Kreegipuu, T. (1987). The mismatch negativity (MMN). In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0081

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, *24*, 375–425. https://doi.org/10.1111/j.1469-8986.1987.tb00311.x

Ngiam, W. X. Q., Adam, K. C. S., Quirk, C., Vogel, E. K., & Awh, E. (2021). Estimating the statistical power to detect set-size effects in contralateral delay activity. *Psychophysiology*, *58*(e13791), 1–10. https://doi.org/10.1111/psyp.13791

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*, 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Olvet, D. M., & Hajcak, G. (2009). The stability of error-related brain activity with increasing trials. *Psychophysiology*, *46*, 957–961. https://doi.org/10.1111/j.1469-8986.2009.00848.x

Pavlov, Y. G., Adamian, N., Appelhoff, S., Arvaneh, M., Benwell, C. S. Y., Beste, C., Bland, A. R., Bradford, D. E., Bublatzky, F., Busch, N. A., Clayson, P. E., Cruse, D., Czeszumski, A., Dreber, A., Dumas, G., Ehinger, B., ... Mushtaq, F. (2021). #EEGManyLabs: Investigating the replicability of influential EEG experiments. *Cortex*, *144*, 213–229. https://doi.org/10.1016/j.cortex.2021.03.013

Pelli, D. G. (1997). The Videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442. https://doi.org/10.1163/156856897x00366

Perez, V. B., & Vogel, E. K. (2012). What ERPs can tell us about working memory. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0180

Picton, T. W. (2010). *Human auditory evoked potentials*. Plural Publishing, Inc.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. R. (2000). Committee report: Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*, 127–152. https://doi.org/10.1111/1469-8986.3720127

Pivik, R. T., Broughton, R. J., Coppola, R., Davidson, R. J., Fox, N., & Nuwer, M. R. (1993). Guidelines for the recording and quantitative analysis of electroencephalographic activity in research contexts. *Psychophysiology*, *30*, 547–558. https://doi.org/10.1111/j.1469-8986.1993.tb02081.x

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*, 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Polich, J. (2012). Neuropsychology of P300. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0089

Pontifex, M. B., Scudder, M. R., Brown, M. L., O'Leary, K. C., Wu, C. T., Themanson, J. R., & Hillman, C. H. (2010). On the number

of trials necessary for stabilization of error-related brain activity across the life span. *Psychophysiology*, *47*, 767–773. https://doi.org/10.1111/j.1469-8986.2010.00974.x

Puce, A., & Hämäläinen, M. S. (2017). A review of issues related to data acquisition and analysis in EEG/MEG studies. *Brain Sciences*, *7*(58), 1–30. https://doi.org/10.3390/brainsci7060058

Rietdijk, W. J., Franken, I. H., & Thurik, A. R. (2014). Internal consistency of event-related potentials associated with cognitive control: N2/P3 and ERN/Pe. *PLoS One*, *9*(e102672), 1–7. https://doi.org/10.1371/journal.pone.0102672

Rossion, B., & Jacques, C. (2012). The N170: Understanding the time course of face perception in the human brain. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0064

Salisbury, D., Collins, K., & McCarley, R. (2010). Reductions in the N1 and P2 auditory event-related potentials in first-hospitalized and chronic schizophrenia. *Schizophrenia Bulletin*, *36*, 99–1000. https://doi.org/10.1093/schbul/sbp003

Sandre, A., Banica, I., Riesel, A., Flake, J., Klawohn, J., & Weinberg, A. (2020). Comparing the effects of different methodological decisions on the error-related negativity and its association with behaviour and gender. *International Journal of Psychophysiology*, *156*, 18–39. https://doi.org/10.1016/j.ijpsycho.2020.06.016

Schimmel, H. (1967). The (±) reference: Accuracy of estimated mean components in average response studies. *Science*, *157*, 92–94. https://doi.org/10.1126/science.157.3784.92

Schupp, H., Cuthbert, B., Bradley, M., Cacioppo, J., Ito, T., & Lang, P. (2000). Affective picture processing: the late positive potential is modulated by motivational relevance. *Psychophysiology*, *37*, 257–261. https://doi.org/10.1111/1469-8986.3720257

Segalowitz, S. J., & Barnes, K. L. (1993). The reliability of ERP components in the auditory oddball paradigm. *Psychophysiology*, *30*, 451–459. https://doi.org/10.1111/j.1469-8986.1993.tb02068.x

Shahin, A., Bosnyak, D., Trainor, L., & Roberts, L. (2003). Enhancement of neuroplastic P2 and N1c auditory evoked potentials in musicians. *The Journal of Neuroscience*, *23*, 5545–5552. https://doi.org/10.1523/jneurosci.23-13-05545.2003

Smaldino, P., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(160384), 1–17. https://doi.org/10.1098/rsos.160384

Smulders, F. T., & Miller, J. (2012). The lateralized readiness potential. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0115

Steele, V. R., Anderson, N. E., Claus, E. D., Bernat, E. M., Rao, V., Assaf, M., Pearlson, G. D., Calhoun, V. D., & Kiehl, K. A. (2016). Neuroimaging measures of error-processing: Extracting reliable signals from event-related potentials and functional magnetic resonance imaging. *NeuroImage*, *132*, 247–260. https://doi.org/10.1016/j.neuroimage.2016.02.046

Steinhauer, K., & Connolly, J. (2008). Event-related potentials in the study of language. In B. Stemmer & H. A. Whitaker (Eds.), *Handbook of the Neuroscience of Language*. Elsevier. https://doi.org/10.1016/b978-0-08-045352-1.00009-4

Stern, J. A., & Davey Smith, G. (2001). Sifting the evidence: What's wrong with significance tests? *British Medical Journal*, *322*, 226–231. https://doi.org/10.1136/bmj.322.7280.226

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language-related ERP components. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0197

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*, 123–138. https://doi.org/10.1111/psyp.12629

Tremblay, K., Ross, B., Inoue, K., McClannahan, K., & Collet, G. (2014). Is the auditory evoked P2 response a biomarker of learning? *Frontiers in Systems Neuroscience*, *8*(28), 1–13. https://doi.org/10.3389/fnsys.2014.00028

Vaughan, H., Costa, L., & Ritter, W. (1968). Topography of the human motor potential. *Electroencephalography and Clinical Neurophysiology*, *25*, 1–10. https://doi.org/10.1016/0013-4694(68)90080-1

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*, 748–751. https://doi.org/10.1038/nature02447

Woods, D. L. (1995). The component structure of the N1 wave of the human auditory evoked potential. *Electroencephalography and Clinical Neurophysiology*, *44*, 102–109.

World Medical Association. (2013). Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Journal of the American Medical Association*, *310*, 2191–2194.

How to cite this article: Hall, L., Dawel, A., Greenwood, L.-M., Monaghan, C., Berryman, K., & Jack, B. N. (2023). Estimating statistical power for ERP studies using the auditory N1, Tb, and P2 components. *Psychophysiology*, *60*, e14363. https://doi.org/10.1111/psyp.14363