Statistical Power Analysis for Event-Related Potential Research: Influencing Factors and Methods

Authors: Nian Jingqing, Chen Xi, Chen Fangfang, Niu Xia, Luo Yu, Niàn Jìngqíng, Luo Yu

Date: 2024-08-24T00:00:00+00:00

Abstract

Statistical power is one of the key indicators for evaluating the robustness and replicability of research findings. However, the standardization and completeness of calculating and reporting statistical power in event-related potential (ERP) research remain to be strengthened. This paper, by reviewing and summarizing the influencing factors, methods, and application examples of statistical power in ERP research, can provide a reference basis for researchers to calculate and report statistical power during the design or preregistration stages of ERP study protocols.

Full Text

Statistical Power Analysis of Event-Related Potential Studies: Influencing Factors and Methods

Nian Jingqing[1], Chen Xi[2], Chen Fangfang[3], Niu Xia[4], Luo Yu[1]*

[1]School of Psychology, Guizhou Normal University, Guiyang 550025, China
[2]OPPO Shanghai Research Institute, Shanghai 200032, China
[3]Department of Clinical Psychology, Wuhu Hospital Beijing Anding Hospital Affiliated to Capital Medical University, Wuhu 241000, China
[4]School of Nursing, Anhui Medical University, Hefei 230031, China

Abstract

Statistical power is one of the key indicators for assessing the robustness and replicability of research results. However, the standardization and completeness

of calculating and reporting statistical power in event-related potential (ERP) studies still need improvement. This paper reviews and summarizes the factors influencing statistical power, methods for calculation, and application examples in ERP studies. It aims to provide researchers with a reference for calculating and reporting statistical power during the design or pre-registration stages of ERP research plans.

Keywords: EEG, event-related potential, statistical power, sample size, number of trials

1. Introduction

Against the backdrop of the reproducibility crisis in psychological research (Nie, Wang, & Luo, 2016; Hu et al., 2016), the robustness and reproducibility of research findings have become crucial for the advancement of psychological science. Statistical power is a key metric for evaluating the reliability and replicability of research results, as it determines the confidence level of findings (Fraley & Vazire, 2014; Schweizer & Furley, 2016). Statistical power refers to the probability of correctly rejecting the null hypothesis when it is false, typically denoted as 1-β and conventionally set at 0.8 (Cohen, 1988). In hypothesis testing, the main parameters of statistical power analysis models include effect size, sample size, Type I error rate (α), and Type II error rate (β). Previous research has thoroughly examined the relationships among these parameters and their connections to statistical power, along with application examples in conventional experimental contexts (Sommet, Weissman, Cheutin, & Elliot, 2023; Peng, Zhang, & Zhou, 2023; Zhai, Li, & Wei, 2022; Hu, 2010; Hu & Dai, 2011, 2017; Zhao & Wang, 2019; Vankelecom, Loeys, & Moerkerke, 2024), which will not be reiterated here. Taking the common independent-samples t-test for continuous variables as an example (calculation details see Appendix 1, available at: https://www.scidb.cn/anonymous/QTd2bVly), when sample size and α level are fixed, statistical power decreases as the effect size (Cohen's d) diminishes (resulting in high β levels). In short, during statistical power analysis, sample size, effect size, Type I error rate (α), and Type II error rate (β) are interrelated functions; when three parameters are determined, the fourth can be calculated through appropriate algorithms, though computational methods vary across different statistical models.

Since Cohen identified the problem of low statistical power in psychological research (Cohen, 1962), increasing attention has been paid to this issue and its consequences, yet the problem remains unresolved. A review of research over the past 60 years from a statistical power perspective reveals that scientific research has an average statistical power of approximately 24% (Smaldino & McElreath, 2016). In neuroscience specifically, statistical power ranges between 8% and 31% (Button et al., 2013), meaning that with a conventional Type I error rate of 5%, the Type II error rate in neuroscience research falls between 69%

and 92%—far exceeding Cohen's recommended 20% threshold and potentially causing researchers to miss many interesting effects (Ioannidis, 2005; Munafò et al., 2017).

EEG technology is one of the most important and widely used tools in cognitive neuroscience. Among EEG studies, event-related potentials (ERPs) have been extensively employed to investigate cognitive processing due to their stable latency and waveform characteristics. However, previous meta-analyses have found that many ERP studies fail to conduct appropriate statistical power analyses, resulting in low statistical power and poor replicability (Clayson et al., 2019). This may be because ERP research involves a more complex multilevel hierarchical model compared to classical statistical power analysis procedures. On one hand, ERP research follows the principle of within-experiment replication. As shown in Figure 1 [Figure 1: see original paper], studies typically require repeated measurements of participants' responses under specific conditions, followed by averaging across multiple trials. This means that for each individual participant, the collected data actually comprise multiple trials. Specifically, researchers can influence statistical power by manipulating parameters such as the number of participants, number of trials, and noise levels. However, in previous ERP studies, statistical power analyses have focused primarily on how many participants to test, largely neglecting how many trials each participant should complete, with trial numbers often going unreported (Larson & Carbine, 2017). Moreover, due to constraints on time and research funding, researchers must often balance participant numbers against trial numbers. This means that even when sample size equals participant count, opaque trial numbers or the use of vague, highly variable heuristics rather than explicit formulas for determining trial numbers (Jensen & MacDonald, 2023; Larson & Carbine, 2017), combined with different statistical analysis methods, can introduce additional measurement error across multiple hierarchical levels, thereby reducing statistical power.

Figure 1 presents a classic schematic of within-subject experimental design in ERP research. The goal is to measure the success rate of specific ERP components across n participants in m conditions, with each participant completing k trials per condition. The noise level for each condition equals the number of trials yielding the ERP component divided by the total number of trials, where outcome 1 indicates observation of the corresponding ERP component and outcome 0 indicates no observation. Two statistical approaches can determine whether success rates differ significantly from chance levels: (1) significance testing of success rates (continuous data) for each participant in each condition, or (2) significance testing of overall success rates (discrete data). The figure illustrates that beyond participant numbers, statistical power may also be influenced by noise levels, trial numbers per participant per condition, and statistical analysis methods.

On the other hand, compared to relatively mature unidimensional data such as scale means or reaction times, EEG data represent a special type of multidi-

mensional time series data with systematic relationships across frequency, time, and voltage amplitude dimensions. These relationships have given rise to various analysis techniques including time-domain analysis, spectral analysis, and time-frequency analysis (Zhao, Li, Chen, & Lei, 2020). Consequently, similar to ambiguous trial numbers, researchers' experimental protocols (variables of interest, experimental design, component differences), equipment factors (number of channels, acquisition protocols), and preprocessing decisions (analysis techniques, signal processing and feature extraction, variable selection) can introduce additional error. This opacity similarly affects statistical power through measurement error, making traditional statistical power analysis methods difficult to apply accurately.

In summary, conducting statistical power analysis in ERP research from a multilevel hierarchical model perspective requires consideration of numerous input parameters, and reconstructing the intrinsic relationships among these parameters into computational procedures remains challenging. Existing research demonstrates that comprehensively considering factors influencing statistical power analysis in ERP studies (e.g., participant numbers, trial numbers) and conducting a priori analyses can help ensure appropriate statistical power and robust experimental results, thereby mitigating the reproducibility crisis (Clayson, Carbine, Baldwin, & Larson, 2019). Furthermore, with the implementation of pre-registration systems, researchers must explicitly plan and provide adequate justification for design elements affecting statistical power, such as participant and trial numbers, in their pre-registration reports (Paul, Govaart, & Schettino, 2021; Zhao, Xia, & Hu, 2024). Therefore, this study reviews and summarizes the influencing factors, methods, and application examples of statistical power analysis in ERP research to provide researchers with reference materials for calculating and reporting statistical power during study design and/or pre-registration.

2. Factors Influencing Statistical Power Analysis in ERP Research

Statistical power analysis in ERP research must consider at least three levels: experimental protocol, implementation/data quality control (measurement precision), and data analysis. Experimental protocol includes characteristics of the ERP component of interest/researcher variables, participant numbers, trial numbers, study design (within-subject, between-subject, mixed designs), research paradigm, and expected ERP effect magnitude. Implementation/data quality control includes equipment factors (number of channels, acquisition protocol) and environmental noise control. Data analysis includes analysis techniques (time-domain analysis methods), signal processing and feature extraction, and statistical analysis methods. A review of existing literature reveals that researchers currently focus primarily on four factors affecting ERP statistical power analysis: participant numbers, trial numbers, effect magnitude, and study design (Boudewyn, Luck, Farrens, & Kappenman, 2018; Gibney et al.,

2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021).

2.1 Participant Numbers

Participant numbers refer to the number of subjects in a study. As a core parameter in statistical power analysis models, increasing participant numbers significantly enhances statistical power. In ERP research, small sample sizes are a direct cause of low statistical power. When conducting statistical power analysis, increasing participant numbers improves power more substantially than increasing trial numbers (Gibney et al., 2020). For example, Gibney et al. (2020) found that in between-subject designs with only 10 participants per group, the likelihood of obtaining truly significant results is extremely low.

2.2 Trial Numbers

Trial numbers represent the relatively small number of repeated measurements needed to collect sufficient data for research purposes. ERPs are relatively small signals within EEG data, and researchers typically extract them by averaging multiple trials of a specific event. Therefore, signal-to-noise ratio (SNR)—the ratio of signal level to noise level in EEG data—is a crucial factor affecting ERP statistical power (Clayson et al., 2013; Kim et al., 2023; W. Zhang & Kappenman, 2024), and SNR improves with the square root of the number of trials averaged (Boudewyn et al., 2018). Specifically, when other conditions are equal, more trials used for averaging yield higher SNR, thereby increasing effect size and statistical power. Research has shown that when participant numbers are small and effect sizes are moderate, approximately doubling trial numbers can effectively increase statistical power to appropriate levels (Boudewyn et al., 2018).

2.3 Effect Magnitude

Effect magnitude refers to the absolute value of an effect measured in microvolts. Specifically, effect magnitude (V) = |ERP component mean amplitude in Condition A − ERP component mean amplitude in Condition B|. In ERP research, effect size typically represents the difference in ERP amplitudes across conditions (effect magnitude). Studies indicate that effect magnitude is inversely related to required trial numbers; ERP components with larger effect magnitudes often achieve stable statistical power with fewer trials (Baker et al., 2021; Boudewyn et al., 2018). For example, in within-subject designs, when the effect magnitude between conditions is large, changes in participant or trial numbers have minimal impact on statistical power. When effect magnitude is moderate, variations in participant and trial numbers substantially affect statistical power. Additionally, if trial numbers are sufficiently large but ERP component effect magnitude is small, appropriate statistical power can still be achieved by increasing participant numbers.

2.4 Study Design

Study design refers to the experimental treatment plan (e.g., within-subject, between-subject, mixed designs) and associated statistical analyses (e.g., t-tests, ANOVA, linear model analysis). Specifically, researchers must clearly define experimental treatment levels before conducting a study, with more treatment levels generally requiring more participants and trials. As shown in Figure 2 [Figure 2: see original paper], with equivalent effect magnitudes, changes in statistical power in within-subject designs depend on trial numbers, whereas in between-subject designs they depend on participant numbers. In other words, under the same effect magnitude, within-subject designs require fewer trials to achieve stable statistical power. For instance, in data simulation studies of within-subject designs, doubling trial numbers can increase statistical power by at least twofold, while doubling participant numbers has a smaller impact (Jensen & MacDonald, 2023).

Figure 2 presents partial key results from simulations of within-subject and between-subject experimental designs. In within-subject designs, trial numbers have a more significant impact on statistical power (++). In between-subject designs, participant numbers have a more significant impact (++). When effect magnitude exhibits floor or ceiling effects (dashed lines)—that is, when effect magnitude is too large or too small—increasing participant and/or trial numbers has minimal impact on statistical power. Figure adapted from Jensen & MacDonald, 2023.

3. Methods and Application Examples for Statistical Power Analysis in ERP Research

Statistical power analysis is primarily based on Null Hypothesis Significance Testing (NHST), calculating different combinations of core parameters to achieve predetermined power levels (Liu et al., 2024). In empirical research, scientifically and rationally planning sample size is a core component of statistical power analysis, given constraints on time and funding (Lakens, 2022). Therefore, when planning sample sizes for ERP studies, researchers must adopt a multilevel hierarchical model perspective (as shown in Figure 1) to comprehensively consider the interrelationships among participant numbers, trial numbers, effect magnitude, and other influencing factors across different study designs to obtain optimal sample size solutions. To achieve this optimal solution, researchers have employed methods such as Post-Hoc Simulations, Monte Carlo Simulations, and Power Contours Plot to analyze statistical power in ERP research. These methods each have distinct emphases: post-hoc simulations focus on the minimum trial numbers needed to obtain ERP components (Thigpen et al., 2017); Monte Carlo simulations emphasize flexible combinations of parameters including participant numbers, trial numbers, effect magnitude, and study design to generate different statistical power analysis models for subsequent analysis (Boudewyn et al., 2018); power contour plots dynamically

adjust participant and trial numbers while considering measurement precision and sample standard deviation (σs) to achieve appropriate statistical power (Baker et al., 2021). Additionally, applying these methods requires access to pilot EEG data or existing EEG datasets.

3.1 Post-Hoc Simulations

The purpose of post-hoc simulation is to help researchers stably estimate specific ERP components based on existing data. If the goal is to obtain the amplitude of a particular ERP component, this essentially resembles parameter recovery in simple linear models (or statistical models for one-sample t-tests)—that is, determining whether an experimental stimulus elicits a specific ERP waveform and how many repetitions are needed to estimate that waveform stably.

The specific procedure involves: conducting a pilot experiment to obtain the desired ERP component, using the number of trials that yielded robust ERP components as the population (K), then drawing subsamples (k) of EEG data with specific trial numbers from this population, averaging these subsamples, and comparing the resulting ERP components with those from the full sample. This process is repeated until ERP components comparable to the population are obtained in subsamples, thereby determining the minimum trial number required. Similarity between population and subsamples is assessed using indices such as correlation coefficients, internal consistency coefficients (Olvet & Hajcak, 2009; Thigpen, Kappenman, & Keil, 2017), test-retest reliability (Huffmeijer, Bakermans-Kranenburg, Alink, & Van IJzendoorn, 2014; Segalowitz & Barnes, 1993), and equivalence (Marco-Pallares, Cucurell, Münte, Strien, & Rodriguez-Fornells, 2011; Pontifex et al., 2010). For example, using internal consistency coefficients: values above 0.90 indicate excellent consistency, 0.70–0.90 indicate high consistency, 0.50–0.70 indicate moderate consistency, and below 0.50 indicate poor consistency. Thigpen et al. (2017) used internal consistency coefficients as a metric in post-hoc simulations to determine the minimum trial numbers needed to obtain P1, N1, and P3 components. During simulation, they drew subsamples with varying trial numbers (10–80 in increments of 10), averaged them, and compared mean amplitudes and SNR of corresponding components with those from the full sample (approximately 80 trials). Results showed that when subsample trial numbers reached 40 or more, internal consistency coefficients between subsample and full-sample ERP components exceeded 0.8, indicating that at least 40 trials are sufficient to obtain relatively robust P1, N1, and P3 components—80 trials are not necessary.

In application, post-hoc simulations have been used to determine trial numbers for ERP components including error-related negativity (ERN), error positivity (Pe), N100, N200, vertex positive potential (VPP)/N170, mismatch negativity (MMN), feedback-related negativity (FRN), late positive potential (LPP), and P300 (Duncan et al., 2009; Fischer, Klein, & Ullsperger, 2017; Huffmeijer et al., 2014; Jill Cohen & Polich, 1997; Larson, Baldwin, Good, & Fair, 2010; Marco-Pallares et al., 2011; Olvet & Hajcak, 2009; Pontifex et al., 2010; Riet-

dijk, Franken, & Thurik, 2014; Segalowitz & Barnes, 1993; Steele et al., 2016; Thigpen et al., 2017).

Post-hoc simulations provide computational basis for determining the minimum trial numbers needed to obtain robust individual ERP components, thereby reducing time costs in ERP research to some extent. However, this method only addresses trial-level planning and cannot accommodate more complex experimental situations.

3.2 Monte Carlo Simulations

Compared to post-hoc simulations, Monte Carlo simulations follow the conventional statistical power analysis approach—simulating and analyzing effect sizes for specific parameters within a statistical model to obtain power estimates for various combinations of trial, condition, and participant parameters. In ERP research, researchers flexibly define statistical power analysis models by dynamically combining participant numbers, trial numbers, effect magnitude, and study design. The fundamental principle involves specifying a virtual population (distribution) to generate virtual samples. In Monte Carlo simulations for ERP research, researchers use pilot or previously collected EEG data as the specified population, adding artificial effects to obtain true effect magnitudes for within-subject and between-subject analyses (Kiesel, Miller, Jolicøeur, & Brisson, 2008; Smulders, 2010; Ulrich & Miller, 2001).

The basic procedure is: as shown in Figure 1, randomly draw n participants with replacement from the participant sample (N). Then, for each drawn participant, randomly draw m conditions with k trials each from their valid trials. Average the k trials for each condition, then add or subtract between-group or between-condition data to obtain corresponding effect magnitudes. Subsequently, conduct difference tests using appropriate statistical methods (e.g., t-tests). Perform 1,000 simulations for each combination of participant numbers, trial numbers, and effect magnitude, calculating the probability of achieving significance across these 1,000 simulations. For example, Boudewyn et al. (2018) conducted Monte Carlo simulations on the ERN component by having 40 participants complete 400 trials of a Flanker task while EEG data were recorded. Based on these data, they simulated 1,000 datasets using Monte Carlo methods and performed comparative analyses. Results showed that with more than 10 participants, stable statistical power above 0.8 for ERN components could be achieved with only 6 trials. Across different experimental designs, t-test requirements for participant and trial numbers varied significantly depending on effect magnitude. In within-subject designs, to achieve power above 0.8 with 20 participants, only 8 trials were needed when effect magnitude was 4 V, but 16 trials were required when effect magnitude was 2 V. In between-subject designs, to achieve power above 0.8 with 6 trials, only 16 participants were needed when effect magnitude was 7 V, but 32 participants were required when effect magnitude was 5 V.

In application, Monte Carlo simulation analysis has been used for statistical power analysis of ERP components including LRP, ERN, N170, MMN, P3, N2pc, N400, CDA, N1, Tb, and P2 (Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021). To facilitate practical application, Hall et al. (2023) provided an online tool, ERP Power Calculator (available at: https://bradleynjack.shinyapps.io/ErpPowerCalculator/), allowing auditory ERP researchers to calculate statistical power by selecting specific ERP components (N1/Tb/P2), trial numbers (20–1,000), participant numbers (10–100), effect magnitude (0–3 V), experimental design (within/between-subject), and alpha levels (0.05/0.01/0.005/0.001). In visual working memory research, Ngiam et al. (2021) offered the online CDA Power Calculator (available at: https://williamngiam.shinyapps.io/CDAPower/), which enables flexible calculation of relevant metrics by selecting effects of interest (robust CDA component/memory load 2 vs. 4/memory load 2 vs. 6) and adjusting combinations of participant numbers, clean trial numbers, and statistical power. Jensen and MacDonald (2023) publicly shared code resources on the OSF platform (available at: https://osf.io/wv3da/) for simulating statistical power across seven ERP components (LRP, ERN, N170, MMN, P3, N2pc, N400) by dynamically combining participant numbers, trial numbers, effect magnitude, and experimental design parameters.

3.3 Power Contour Plots

As previously discussed, beyond participant numbers, trial numbers, effect magnitude, and experimental design, we must reconsider measurement precision (the mean of true scores/total scores across trials)—a key indicator that influences statistical power through measurement error (Nebe et al., 2023). In this context, measurement precision refers to the ability to obtain similar results when repeatedly measuring a variable with constant true scores (Cumming, 2014), which is related to multiple factors including trial numbers, equipment, and ERP component differences. In ERP research, ERP component latencies and waveforms lack strict consistency and stability, leading to errors across time, individuals, and trials. This increased measurement error reduces statistical power (Nebe et al., 2023).

Baker et al. (2021) proposed power contour plots, which essentially visualize results from Monte Carlo simulations of specific statistical models. Their key feature is the separation of within-subject and between-subject variance, with an underlying hierarchical statistical model combined with various experimental scenarios for Monte Carlo modeling. Specifically, as shown in Figure 3 [Figure 3: see original paper], power contour plots dynamically adjust participant and trial numbers under constraints of within-participant variance (σw) and between-participants variance (σb) to calculate corresponding statistical power until results reach predetermined criteria. Points representing combinations of participant numbers (N) and trial numbers (k) with equal power are connected to form contour lines, with multiple contours representing different

power levels (Baker et al., 2021). In practice, researchers can identify an ideal balance point between participant and trial numbers through power contours, selecting appropriate numbers based on actual conditions while meeting power requirements and minimizing research costs. For example, Baker et al. (2021) resampled participant and trial numbers for P100 and N600 components based on existing EEG data and plotted corresponding power contours. Results showed that for P100 components, statistical power increased with both participant and trial numbers when sample bias was small. For N600 components, statistical power depended heavily on participant numbers, though trial numbers could be increased to reduce required participant numbers when trial numbers were relatively small ($k < 200$).

Figure 3 illustrates a power contour plot schematic (color version available online). Participant numbers (N) range from 0–30, trial numbers (k) from 0–600, α level is 0.05, mean difference is 1.32, within-participant measurement error is 12 V, and between-participants measurement error is 1.1 V. The green point represents the ideal balance between participant and trial numbers when statistical power reaches 80% under these conditions, with $N = 16$ and $k = 79$. The schematic was generated using the online Power Contour Estimation tool developed by Baker et al. (2021).

In application, power contour plots have been used to calculate ideal combinations of participant and trial numbers for ERP components including P100, P200, and N600, as well as for the alpha frequency band (8–12 Hz) (Baker et al., 2021). To facilitate practical use, Baker et al. (2021) developed the online Power Contour Estimation tool (available at: https://shiny.york.ac.uk/powercontours/), which calculates statistical power and ideal participant-trial combinations by inputting parameters including participant numbers, trial numbers, alpha level, mean difference, within-subject standard deviation, between-subject standard deviation, and recruitment costs.

4. Challenges in Statistical Power Analysis for ERP Research

Existing research has systematically examined how participant numbers, trial numbers, effect magnitude, and experimental design interact to influence statistical power (Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021). However, future research should address four additional considerations:

4.1 Attention to Ceiling and Floor Effects

Previous research has shown that statistical power changes with participant and trial numbers, but when ceiling or floor effects occur, variations in these parameters have minimal impact on power (Boudewyn et al., 2018).

4.2 Attention to Signal-to-Noise Ratio Effects

While previous research on measurement precision has focused primarily on trial numbers, measurement precision reductions caused by other factors should not be overlooked. The signal-to-noise ratio emphasized in EEG research—essentially a measurement precision issue—also reduces statistical power. ERP research SNR is influenced by research paradigms (experimental protocols), EEG data acquisition (equipment factors such as different acquisition environments, devices, and impedance levels) (Kappenman & Luck, 2010; Laszlo et al., 2014; Luck & Kappenman, 2017; Puce & Hämäläinen, 2017), signal processing and feature extraction (Clayson et al., 2021; Delorme, 2023; Sandre et al., 2020; G. Zhang, Garrett, & Luck, 2024a, 2024b; G. Zhang, Garrett, Simmons, et al., 2024; G. Zhang & Luck, 2023), and statistical testing methods (Luck & Gaspelin, 2017). However, Monte Carlo simulations cannot effectively model the true SNR level in each EEG dataset. Notably, researchers' subjective or inadvertent decisions in signal-to-variable data transformation pipelines (e.g., different processing and analysis pipelines) may also produce false-positive results (Luck & Gaspelin, 2017). Therefore, these additional factors affecting SNR represent an important direction for future statistical power research.

4.3 Validation in More Complex Experimental Contexts

Existing research has simulated relationships among participant numbers, trial numbers, effect magnitude, and statistical power in within-subject and between-subject designs (Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021). Since EEG data quality varies across paradigms, participants, and measurement indices (G. Zhang & Luck, 2023), whether existing conclusions apply to more complex experimental designs (e.g., mixed designs) and analysis methods (e.g., multifactor designs) requires further investigation.

4.4 Cautious Generalization of Existing Conclusions

Current conclusions derive from data simulations of specific ERP component mean amplitudes, representing relatively ideal results that may not generalize to datasets or analysis methods substantially different from the simulation datasets. For instance, when using component latency as a measurement index or time-frequency analysis methods in ERP research, additional factors beyond amplitude (e.g., phase) may need consideration. Future research should more comprehensively evaluate other potential factors influencing statistical power in ERP studies and develop more broadly applicable statistical power analysis methods and computational tools.

5. Future Directions and Recommendations for Statistical Power Analysis in ERP Research

Amid concerns about the robustness and replicability of ERP research findings, increasing attention is being paid to the negative consequences of low-power studies, with calls for a priori statistical power analysis to mitigate these risks. Statistical power holds significant meaning for both authors and readers of ERP research. Optimizing research designs and reducing investment in low-power studies during the design and/or pre-registration stages requires collaborative efforts from all stakeholders.

5.1 Scientific and Rational Sample Size Planning

Researchers should plan sample sizes appropriately during experimental design. Sample size planning is a core component of statistical power analysis, with general principles well-established in previous literature (Lakens, 2022; Sommet et al., 2023). For ERP research, we recommend using Monte Carlo simulations or power contour plots for sample size planning (Baker et al., 2021; Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021). Additionally, Bayesian factor-based sequential analysis represents an important alternative approach beyond a priori sample size planning (Zheng & Hu, 2023).

5.2 Accurate and Comprehensive Reporting with Cautious Generalization

Researchers must recognize the reproducibility issues in EEG research, particularly ERP studies. Given the complex experimental conditions, equipment, and measurement modalities in cognitive neuroscience, researchers should comprehensively report all known experimental conditions and parameters to provide essential metadata for reproducibility and power analysis (Luo, Nian, & Wang, 2021). Simultaneously, researchers must acknowledge study limitations and report conclusions cautiously, especially regarding generalizability.

5.3 Adoption of Established, Peer-Recognized Protocols

Researchers should recognize the importance of evidence-based practices. Any modifications to previous studies (e.g., regions of interest and channel locations) must be justified with solid evidence during literature review (Dien, 2017), avoiding post-hoc, data-driven analyses.

We sincerely thank the anonymous reviewers for their guidance and assistance. We are grateful to the Chinese Open Science Network (COSN) for providing the COSN Summer Hackathon 2023, which enabled the collaboration for this paper. We also thank Professor Hu Chuanpeng from Nanjing Normal University, postdoctoral researcher Zhang Guanghui from Professor Steven Luck's lab at UC Davis, Wu Rui from Guizhou University of Traditional Chinese Medicine, Zhang

Huoyin from Shenzhen University, Gan Yetong from University of Science and Technology of China, Sun Mingze and Wen Xiujuan from South China Normal University, and Guo Rui and Yang Jiyue from Southwest University for their help and suggestions during manuscript preparation.

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(3), 295–314.

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(6), e13049.

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.

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.

Clayson, P. E., Baldwin, S. A., 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.

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(11), e13437.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65(3), 145.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Cohen, J., & Polich, J. (1997). On the number of trials needed for P300. *International Journal of Psychophysiology*, 25(3), 249–255.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29.

Delorme, A. (2023). EEG is better left alone. *Scientific Reports*, 13(1), 2372.

Dien, J. (2017). Best practices for repeated measures ANOVAs of ERP data: Reference, regional channels, and robust ANOVAs. *International Journal of Psychophysiology*, 111, 42–56.

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.

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(7), 998–1009.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, 9(10), e109019.

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(2), e13482.

Hall, L., Dawel, A., Greenwood, L., Monaghan, C., Berryman, K., & Jack, B. N. (2023). Estimating statistical power for ERP studies using the auditory N1, tb, and P2 components. *Psychophysiology*, e14363.

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, 160–165.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

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*, 60(7), e14245.

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology*, 47(5), 888–904.

Kiesel, A., Miller, J., Jolicøeur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, 45(2), 250–274.

Kim, B., Erickson, B. A., Fernandez-Nunez, G., Rich, R., Mentzelopoulos, G., Vitale, F., & Medaglia, J. D. (2023). EEG phase can be predicted with similar accuracy across cognitive states after accounting for power and signal-to-noise ratio. *eNeuro*, 10(9).

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), 33267.

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(6), 1160–1165.

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.

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.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, 54(1), 146–157.

Luck, S. J., & Kappenman, E. S. (2017). Electroencephalography and event-related brain potentials. In *Handbook of psychophysiology* (pp. 74–100). Cambridge University Press.

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(6), 852–860.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 1–9.

Nebe, S., Reutter, M., Baker, D. H., Bölte, J., Domes, G., Gamer, M., … Feld, G. B. (2023). Enhancing precision in human neuroscience. *eLife*, 12, e85980.

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(5).

Olvet, D. M., & Hajcak, G. (2009). The stability of error-related brain activity with increasing trials. *Psychophysiology*, 46(5), 957–961.

Paul, M., Govaart, G. H., & Schettino, A. (2021). Making ERP research more transparent: Guidelines for preregistration. *International Journal of Psychophysiology*, 164, 52–63.

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(4), 767–773.

Puce, A., & Hämäläinen, M. S. (2017). A review of issues related to data acquisition and analysis in EEG/MEG studies. *Brain Sciences*, 7(6), 58.

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(7), e102672.

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.

Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise*, 23, 114–122.

Segalowitz, S. J., & Barnes, K. L. (1993). The reliability of ERP components in the auditory oddball paradigm. *Psychophysiology*, 30(5), 451–459.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9), 160384.

Smulders, F. T. (2010). Simplifying jackknifing of ERPs and getting more out of it: Retrieving estimates of participants' latencies. *Psychophysiology*, 47(2), 387–392.

Sommet, N., Weissman, D. L., Cheutin, N., & Elliot, A. J. (2023). How many participants do I need to test an interaction? Conducting an appropriate power analysis and achieving sufficient power to detect an interaction. *Advances in Methods and Practices in Psychological Science*, 6(3), 25152459231178728.

Steele, V. R., Anderson, N. E., Claus, E. D., Bernat, E. M., Rao, V., Assaf, M., … 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.

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, 54(1), 123–138.

Ulrich, R., & Miller, J. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology*, 38(5), 816–827.

Vankelecom, L., Loeys, T., & Moerkerke, B. (2024). How to safely reassess variability and adapt sample size? A primer for the independent samples t test. *Advances in Methods and Practices in Psychological Science*, 7(1), 25152459231212128.

Zhang, G., Garrett, D. R., & Luck, S. J. (2024a). Optimal filters for ERP research I: A general approach for selecting filter settings. *Psychophysiology*, e14531.

Zhang, G., Garrett, D. R., & Luck, S. J. (2024b). Optimal filters for ERP

research II: Recommended settings for seven common ERP components. *Psychophysiology*, e14530.

Zhang, G., Garrett, D. R., Simmons, A. M., Kiat, J. E., & Luck, S. J. (2024). Evaluating the effectiveness of artifact correction and rejection in event-related potential research. *Psychophysiology*, 61(5), e14511.

Zhang, G., & Luck, S. J. (2023). Variations in ERP data quality across paradigms, participants, and scoring procedures. *Psychophysiology*, 60(7), e14264.

Zhang, W., & Kappenman, E. S. (2024). Maximizing signal-to-noise ratio and statistical power in ERP measurement: Single sites versus multi-site average clusters. *Psychophysiology*, 61(2), e14440.

Appendix 1: Statistical Power Analysis—Example of Common Independent-Samples Two-Tailed t-Test (Continuous Variables)

As shown in Appendix Figure 1, the black area represents β (Type II error rate), the red area represents α (Type I error rate), and the t-distribution critical value (t_{crit}) marks the cutoff point for the red area. Note that when $\alpha = 0.05$, the cutoff points represent 2.5% of the area in each tail of the distribution represented by 1. The shape of the t-distribution is determined by degrees of freedom (df) and non-centrality parameter (δ), so sample size also affects t_{crit}.

Appendix Figure 1. Schematic diagram of statistical power and significance testing. Sample size (n) = 20, significance level (α) = 0.05, statistical power (1-β) = 0.8, effect size (Cohen's d) = 0.63. Diagram generated using the interactive web tool developed by Kristoffer Magnusson (https://rpsychologist.com/d3/nhst/).

In a model with parameter values , the statistical power function of the testing procedure is P_ (Reject), where represents the parameters of the statistical test and P_ (Reject) represents the probability of rejecting the null hypothesis (H0). For a two-tailed independent-samples t-test, H0 assumes no significant difference between population means of two independent samples (H0: 1 = 2), while the alternative hypothesis posits a significant difference (H_a/2: 1 ≠ 2, two-tailed). As shown in Appendix Figure 1, P_ (Reject) corresponds to the right portion of the alternative hypothesis distribution cut off by t_{crit}, with the left black portion representing β. Since the total probability density function integrates to 1, the right portion equals 1-β, i.e., P_ (Reject) = 1-β.

Appendix Figure 1 shows that calculating P_ (Reject) involves multiple parameters: t critical value (t_{crit}), effect size (Cohen's d), and Type I error rate

$\alpha/2$ (two-tailed test). Using post-hoc power calculation as an example, where sample size is determined, the calculation proceeds as follows:

Step 1: Determine effect size d = | 1 - 2| / σ from existing data, where 1 and 2 are sample means and σ is the pooled standard deviation: σ = √[((n1-1)s1^2 + (n2-1)s2^2) / (n1 + n2 - 2)].

Step 2: Calculate degrees of freedom: df = n1 + n2 - 2.

Step 3: Using tables or software, determine the critical value t_{crit} from the central t-distribution based on significance level α and degrees of freedom df. In some cases (non-central t-distribution), the non-centrality parameter δ = d√(n1n2/(n1+n2)) is also required. For two-tailed tests, the critical value is t_{crit} = t(1 - $\alpha/2$, df).

Step 4: Calculate cumulative probabilities for both tails. For the central t-distribution, using the cumulative distribution function F: - Left tail: P(T < -t_{crit} | df) = F(-t_{crit} | df) - Right tail: P(T > t_{crit} | df) = 1 - F(t_{crit} | df)

Therefore, statistical power = 1 - [F(t_{crit} | df) - F(-t_{crit} | df)].

In summary, post-hoc statistical power is determined by α and df. For non-central t-distributions, the non-centrality parameter δ is also required. For different statistical methods, the general conclusions are:

1. **Effect size:** Effect size represents the difference between distributions. Larger effect sizes indicate greater actual differences between distributions. With fixed sample size and Type I error rate α, t_{crit} remains unchanged while statistical power increases. For non-central t-distributions, the non-centrality parameter must also be considered, but the conclusion remains unchanged.

2. **Sample size:** Sample size affects standard error. As sample size increases, pooled standard deviation estimates become more precise and degrees of freedom df increase, causing t_{crit} to decrease and thereby increasing statistical power. With other parameters fixed, larger sample sizes improve the precision of effect size estimates.

3. **Type I error rate α:** α determines the position of t_{crit}. Larger α values result in smaller β values and greater statistical power. In practice, α is typically fixed at 0.05. Therefore, except in studies with predetermined α/β ratios, only sample size, effect size, and statistical power are typically adjustable.

4. **t critical value (t_{crit}):** t_{crit} is jointly determined by Type I error rate α and degrees of freedom df, serving as the cutoff point for rejecting the null hypothesis in hypothesis testing.

Note: Figure translations are in progress. See original paper for figures.