Sample size calculations in human electrophysiology (EEG and ERP) studies: A systematic review and recommendations for increased rigor

Michael J. Larson, Kaylie A. Carbine

ABSTRACT

There is increasing focus across scientific fields on adequate sample sizes to ensure non-biased and reproducible effects. Very few studies, however, report sample size calculations or even the information needed to accurately calculate sample sizes for grants and future research. We systematically reviewed 100 randomly selected clinical human electrophysiology studies from six high impact journals that frequently publish electroencephalography (EEG) and event-related potential (ERP) research to determine the proportion of studies that reported sample size calculations, as well as the proportion of studies reporting the necessary components to complete such calculations. Studies were coded by the two authors blinded to the other's results. Interrater reliability was 100% for the sample size calculations and kappa above 0.82 for all other variables. Zero of the 100 studies (0%) reported sample size calculations. 77% utilized repeated-measures designs, yet zero studies (0%) reported the necessary variances and correlations among repeated measures to accurately calculate future sample sizes. Most studies (93%) reported study statistical values (e.g., F or t values). Only 40% reported effect sizes, 56% reported mean values, and 47% reported indices of variance (e.g., standard deviations/standard errors). Absence of such information hinders accurate determination of sample sizes for study design, grant applications, and meta-analyses of research and whether studies were adequately powered to detect effects of interest. Increased focus on sample size calculations, utilization of registered reports, and presenting information detailing sample size calculations and statistics for future researchers are needed and will increase sample size-related scientific rigor in human electrophysiology research.

© 2016 Elsevier B.V. All rights reserved.

Keywords:
Sample size
Rigor
Event-related potential
ERP
Electroencephalogram
EEG

1. Introduction

Reproducibility and finding “true” effects represent the bedrock of rigorous science. Recently, a focus on clear and transparent replications of previous research (e.g., Open Science Collaboration, 2015), increasing publication of null results (e.g., Cohen, in press), and pre-registering studies (e.g., Baldwin, in press) have been recommended, among other possibilities, to improve the rigor of scientific research. An additional recommendation that may strengthen research is to complete a priori sample size calculations to ensure that studies are sufficiently powered, to reduce the probability of false negative findings, and to ensure that effect sizes are not inflated (Button et al., 2013; Ioannidis, 2005, 2008; Sawyer and Ball, 1981; Sterne and Smith, 2001). Despite the importance of adequate sample sizes, literature reviews continue to show that underpowered studies persist across multiple disciplines (Bezeau and Graves, 2001; Cashen and Geiger, 2004; Chan and Altman, 2005; Maggad et al., 2003). Consequently, scientific research in fields such as neuroscience (Button et al., 2013) and psychology (Open Science Collaboration, 2015) have come under scrutiny for poor reproducibility and low sample sizes. Indeed, some have suggested that low statistical power due to small sample sizes is a pervasive problem that undermines conclusions in much of the neuroscience literature (Button et al., 2013). In this regard, Button et al. (2013) present convincing data that with a small sample size comes a lower probability of replication, exaggerated estimates of effects when a statistically significant finding is reported, and poor positive predictive power of small sample effects.

There are multiple reasons why studies may have small sample sizes and consequently low power. For example, researchers may be limited in their resources to conduct larger studies (Elms, 1975; Nosek et al., 2012). In addition, pressure to quickly publish papers and advance a research career may influence data collection (Nosek et al., 2012), including stopping data collection once a desirable result is obtained (John et al., 2012; Simmons et al., 2011), or conducting multiple low-powered projects instead of fewer high-powered ones (Nosek et al., 2012). Maxwell (2004) also points out that most studies involve multiple hypothesis tests. Even though the power of any single test may be low by a reasonable standard, conducting multiple tests with a lower power makes it more likely that something of interest will emerge as statistically significant, possibly by chance. Furthermore, the lack of
information in the literature, such as effect sizes or variance components, that is needed to perform power analyses may make it difficult to conduct sample size calculations to ensure an appropriate sample size is collected (Guo et al., 2014).

While we sympathize with these influences and have not consistently reported sample size calculations in our own work, adequate power and sample sizes are critical to obtain accurate results and move scientific research forward. Maxwell (2004) shows that the consequence of underpowered studies is the abundance of apparent contradictions in the published literature. For example, in a review that highlighted white matter connectivity and diffusion tensor imaging in autism spectrum disorders (ASD), Ameis and Catani (2015) noted inconsistencies as to whether children/adolescents with ASD exhibited decreased white matter integrity (as measured by decreased fractional anisotropy) or increased white matter integrity (as measured by increased fractional anisotropy). The authors concluded that the relatively small sample sizes (average 20 subjects) in these studies are likely responsible for the inconsistent results. Other authors have similarly shown the importance of adequate power for the development of a cumulative science. For example, Sawyer and Ball (1981) illustrate how replication studies may also fail to detect true effects if researchers do not increase their power based on previous research. Ioannidis (2005) demonstrates how low power and effect sizes along with other issues such as flexibility in design may increase the number of false finding in the literature. Therefore, increased scrutiny on sampling (both sample sizes and variance components) in neuroscience and psychophysiological research may be an avenue to help identify “true” and reproducible effects. One way to address the scrutiny that has come along with small sample sizes is to consistently consider and report sample size calculations used when designing a study. Sample size calculations, such as a power analysis, estimating the number of participants to obtain a desired confidence interval width, or utilizing Bayesian estimators (Lenth, 2001; Wang et al., 2005) inform reviewers and readers if the reported sample size is adequate enough to accurately detect the effects of interest and determine if the effects may be inflated (see Button et al., 2013; Ioannidis, 2005). However, in order to determine an adequate sample size for a desired hypothesis test or analysis, additional parameters such as variance components and effect sizes are needed (Lenth, 2001; Wilkinson, 1999), which are typically obtained from previous research studies, from publicly available data in repositories, or directly from study authors. Therefore, it is imperative that authors report effect sizes and variance components or post data in accessible repositories so that future studies can accurately conduct power analyses in order to have appropriate sample sizes.

Evidence from neuroimaging and clinical studies suggests that quantitative sample size analyses are rarely included in published research. Guo et al. (2014) found that out of 100 randomly selected clinical fMRI papers from six leading journals, only one paper reported a sample size calculation and few reported the statistics needed to conduct future power analyses. In addition, a review of randomized controlled trials found that even when sample size calculations were reported, the reported statistics often lacked all the necessary parameters needed for sample size calculations or were even erroneous (Charles et al., 2009).

Whereas IMRI research perhaps bears much of the inquiry for questions of scientific rigor due to the high visibility of IMRI research, human electrophysiology research, including electroencephalogram (EEG) and event-related potential (ERP) studies, may be equally susceptible to questions and criticisms of reproducibility and small sample sizes. For example, although there are published guidelines on how to conduct and report ERP research (Keil et al., 2014; Picton et al., 2000), researchers may choose to perform ERP analyses in multiple different ways (i.e., peak amplitude vs mean amplitude or using different clusters of electrodes; see Clayson et al., 2013). The variety of options may make it difficult to compare results and replicate studies, particularly if the methods are not clearly reported (Barch and Yarkoni, 2013).

In addition, small sample sizes in clinical ERP research may contribute to discrepant findings in the literature. For example, research on the error-related negativity (ERN), an ERP that reflects performance monitoring, is rather inconsistent in individuals with major depressive disorder (MDD: Moran et al., in press; Vaidyanathan et al., 2012). The discrepancy in findings between studies may be due to differences in MDD severity as individuals with mild to moderate MDD tend to have enhanced ERN amplitudes relative to controls (Chiu and Deldin, 2007; Holmes and Pizzagalli, 2008, 2010), whereas individuals with severe MDD seem to have ERN amplitudes smaller or similar to controls (Georgiadi et al., 2011; Olvet et al., 2010; Ruchswol et al., 2004; Schrijvers et al., 2008; Schrijvers et al., 2009). However, the sample sizes in these studies were relatively small, ranging from 10 to 26 subjects per group, and no sample size calculations were reported in any of the papers. As such, it is possible that the discrepancies are partially due to low power from smaller sample sizes (among other possibilities including reliability of measures, type of depression, heterogeneity in samples and measures, etc.). Performing sample size calculations in clinical EEG and ERP research to ensure there is adequate power to detect the effects of interest may be one area that could contribute in helping to clarify the contradictions seen in the literature.

Despite the importance of quantitatively considering sample size as outlined above, sample size calculations and considerations of statistical power may be infrequently presented and considered in ERP research methodology. In fact, our own lab and research is subject to this critique; however, no studies that we are aware of have systematically examined human electrophysiology (including EEG and ERPs) studies to report if quantitative sample size comparisons are being utilized or the effect sizes and variance components needed to conduct sample size calculation are being reported.

One possible reason for this is that authors are unsure of the procedures for conducting sample size calculations or cannot find the necessary information to report such calculations. Information necessary to complete a sample size estimate varies based on the type of study design and analysis you are completing (e.g., experimental versus correlational designs; longitudinal versus cross-sectional research; regression versus ANOVA versus multi-level modeling, etc.). Several books and articles provide specific formulas and information to assist in determining what methods are most appropriate for determining a specific sample size calculation (Aberson, 2010; Cohen, 1988; Faul et al., 2007; Kim and Seo, 2013; Kirby et al., 2002; Lenth, 2001; Whitley and Ball, 2002; Zhong, 2009). That said, if a sample size calculation is completed from a power-based approach there are some aspects that are needed. Specifically, Lenth (2001) presented five types of information needed for practical sample size determination including: 1) the hypothesis test of interest on the parameter and the probability data (e.g., What is the hypothesis? How many groups are present? What statistical test is most appropriate to test this hypothesis? What is the base-rate/probability of the effect? Etc.); 2) the Type I error rate (α) of the test—traditionally at 0.05; 3) the effect size of a meaningful change or difference in the variable(s) of interest; 4) historical precedent, values, or estimates to determine the power function for the test; 5) the target value for power (1 – Type II error rate [β]). Power of 0.80 would indicate that one in five times a difference that is present in the data is not found. Depending on the statistic, authors may also need the number of predictor variables or levels in the analysis, the expected variances of the data, and the correlations among repeated measurements among other types of information (Guo et al., 2013; Kim and Seo, 2013; Lenth, 2001). With these types of information authors can work toward determining sample size calculations. Yet, it is our experience that sample size calculations occur infrequently in psychophysiological research.

The purpose of the current study was to examine the frequency with which studies include information on sample size calculations in clinical electrophysiological research as well as see if the necessary variance components are present for future sample size calculations. Our
دریافت فوری متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات