Methodological Reporting Behavior, Sample Sizes, and Statistical Power in Studies of Event-Related Potentials: Barriers to Reproducibility and Replicability

Peter E. Clayson*[1,2], Kaylie A. Carbine[3], Scott A. Baldwin[3], and Michael J. Larson[3,4]

[1]Veterans Affairs Greater Los Angeles Healthcare System, Los Angeles, CA

[2]Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA

[3]Department of Psychology, Brigham Young University, Provo, UT

[4]Neuroscience Center, Brigham Young University, Provo, UT

**Corresponding author at*: VA Greater Los Angeles Healthcare System, MIRECC 210A, Bldg. 210, 11301 Wilshire Blvd., Los Angeles, CA 90073, United States. *Email address*: pclayson@ucla.edu

Abstract

Methodological reporting guidelines for studies of event-related potentials (ERPs) were updated in *Psychophysiology* in 2014. These guidelines facilitate the communication of key methodological parameters (e.g., preprocessing steps). Failing to report key parameters represents a barrier to replication efforts, and difficultly with replicability increases in the presence of small sample sizes and low statistical power. We assessed whether guidelines are followed and estimated the average sample size and power in recent research. Reporting behavior, sample sizes, and statistical designs were coded for 150 randomly-sampled articles from five high-impact journals that frequently publish ERP research from 2011 to 2017. An average of 63% of guidelines were reported, and reporting behavior was similar across journals, suggesting that gaps in reporting is a shortcoming of the field rather than any specific journal. Publication of the guidelines paper had no impact on reporting behavior, suggesting that editors and peer reviewers are not enforcing these recommendations. The average sample size per group was 21. Statistical power was conservatively estimated as .72-.98 for a large effect size, .35-.73 for a medium effect, and .10-.18 for a small effect. These findings indicate that failing to report key guidelines is ubiquitous and that ERP studies are primarily powered to detect large effects. Such low power and insufficient following of reporting guidelines represent substantial barriers to replication efforts. The methodological transparency and replicability of studies can be improved by the open sharing of processing code and experimental tasks and by *a priori* sample size calculations to ensure adequately powered studies.

Key Words: event-related potentials (ERPs); reporting guidelines; statistical power; sample size; replicability

Replication difficulties in psychological science have focused attention on research practices that contribute to replication failures (Chambers, 2017; De Boeck & Jeon, 2018; Forstmeier, Wagenmakers, & Parker, 2017; Nelson, Simmons, & Simonsohn, 2018; Shrout & Rodgers, 2018). A common target of criticism is flexible data analysis that inflates the chance of erroneously observing significant effects. Unfortunately, this practice is endemic in cognitive neuroscience and psychophysiology, such as in studies of event-related brain potentials (ERPs; Baldwin, 2017; Larson & Carbine, 2017; Luck & Gaspelin, 2017), because collecting and analyzing ERPs is computationally intensive and requires many methodological choices. Given the numerous possible researcher degrees of freedom in ERP studies, publication guidelines for ERP studies were published in *Psychophysiology* by a committee convened by the Society for Psychophysiological Research (SPR; Keil et al., 2014). These guidelines help to facilitate methodological transparency by identifying the key parameters that should be reported in ERP studies. Such information is critical for evaluating the quality of research and for ensuring sufficient information is present to conduct replication studies. The first purpose of the present study was to evaluate the extent to which these publication guidelines are followed and to determine how the publication of these guidelines influenced reporting behavior in ERP research.

Adhering to systematic publication guidelines for ERP studies ensures that key details are reported and sheds light on critical data processing steps. Ambiguous or missing experimental details hinder replication efforts and can elevate false positive rates in the face of undisclosed flexibility (Carp, 2013). Hence, a purpose of a Methods section is to clearly communicate all of these steps and to provide justification of the relevant decisions in the data processing pipeline. For example, a recent study conducted an informal analysis on whether ERP studies justified the

measurement windows and sites used for ERP analysis (Luck & Gaspelin, 2017), which is a requirement of the publication guidelines (Keil et al., 2014). Four of the fourteen studies (29%) published in *Psychophysiology* failed to sufficiently justify measurement windows and sites. This small analysis indicated that only a minority of studies failed to report this critical step. However, the sample was small and the extent to which all necessary data processing steps are reported for entire studies remains unclear.

Another important research practice associated with replication failures is the use of small sample sizes, an issue that is exacerbated when researcher degrees of freedom are exploited (Forstmeier et al., 2017). Studies of small samples can lead to the attenuation or exaggeration of effect sizes (i.e., magnitude error) or flip the direction of the relationship between variables (i.e., sign error; Brand & Bradley, 2016; Gelman, 2018; Gelman & Carlin, 2014; Loken & Gelman, 2017). Hence, significant findings based on small samples can lead to erroneous statistical inferences that fail to replicate. The bias to believe such findings in small samples is referred to as the "law of small numbers fallacy" (Tversky & Kahneman, 1971). This fallacy reflects the belief that because it is more difficult to observe statistical significance in a small sample than it is to observe significance in a large sample, then finding a statistically significant effect in a small sample must be "true" and represent a robust and real effect. However, in studies of small samples, magnitude and sign errors are common because of the noisy nature of the data. Thus, small sample sizes can undermine the replicability and generalizability of scientific research.

Sample sizes are an important determinant of statistical power, and statistical power is impacted by multiple other factors, such as the statistical analysis approach, number of observations/trials, effect size, and reliability of measurements. Low statistical power is

prevalent in neuroscience studies (Button et al., 2013b) and has been observed for some ERP research (Baldwin, 2017). In a meta-analysis of the relationship between the error-related negativity (ERN) and anxiety, the average number of participants per group was 22 (Moser, Moran, Kneip, Schroder, & Larson, 2016). The average statistical power of these studies was conservatively estimated at around 10% to detect the estimated effect size of Cohen's d = -.36 (Baldwin, 2017). According to the ERP guidelines paper, statistical power and/or effect sizes should be explicitly stated when they are relevant to the research question (Keil et al., 2014). However, a systematic review of 100 clinical electroencephalogram (EEG) and ERP studies recently found that only 40% of studies reported effect sizes and no study (0%) reported *a priori* power analyses, suggesting that statistical power is rarely reported in ERP research (Larson & Carbine, 2017).

Low statistical power is also associated with the exaggeration of significant effects in the presence of researcher bias (Button et al., 2013a; Ioannidis, 2005; Ioannidis, 2008). For example, exploiting researcher degrees of freedom and conducting many different tests of effects increases the likelihood of finding a spuriously large effect, because only large statistical effects will satisfy a statistical significance threshold in a study with low power (Ioannidis, 2005; Ioannidis, 2008). Such researcher degrees of freedom are reflected in some ERP analysis approaches. For example, if a researcher fails to find a significant effect at the *a priori* electrode of interest, it is possible to look at additional electrodes, such as all sites along the midline or multiple lateralized sites. In this way, ERP studies can reduce statistical power but inflate chances to find significant effects by examining multiple channels (Baldwin, 2017; Luck & Gaspelin, 2017). Taking multiple analysis approaches until a significant finding is obtained potentially leads to erroneous conclusions, but this practice can be difficult to detect when such exploratory analyses are

presented as confirmatory. Because of the relationship between statistical power and the statistical analysis performed, it is important to consider the statistical analysis approaches used in ERP studies to estimate the power in the literature.

The present study focused on reporting behavior and sample sizes of ERP studies, which are two important aspects for replication. The first aim was to determine the extent to which ERP studies published in multiple journals that focus on ERP research follow the publication guidelines in *Psychophysiology* (Keil et al., 2014). Then, we examined whether the publication of the guidelines paper impacted reporting behavior by testing differences in guideline reporting pre- and post-guideline publication. We also sought to determine the typical sample sizes of ERP studies; the average statistical power of the ERP literature was then computed based on the typical sample size and statistical designs.

Method

The aims, inclusion/exclusion criteria, design, and analyses were preregistered on Open Science Framework (OSF; https://osf.io/pdbw3/). Raw data, source code for all analyses, and supplementary material are also posted on OSF (https://osf.io/mbsvy/).

Article Selection

To determine the journals to use for coding ERP studies, the top 10 journals with the most ERP studies published between 2011 and 2013 and between 2015 and 2017 were identified from PubMed. Journals were then ranked based on their 2017 impact factors. The top three journals were selected and included *NeuroImage*, *Clinical Neurophysiology*, and *Journal of Cognitive Neuroscience*. We also included the two flagship society journals for ERP research, *Psychophysiology* and *International Journal of Psychophysiology*.

The ERP guidelines paper was published in 2014 (Keil et al., 2014). Fifteen ERP studies were randomly sampled without replacement from each abovementioned journal for the following years: 2011, 2012, 2013, 2015, 2016, and 2017. Articles from 2014 were not examined, because it was expected that there would be a delay in the adoption of the guidelines. The first five articles sampled from each journal from each year that satisfied the inclusion and exclusion criteria were selected for subsequent coding. This resulted in a total of 150 articles. Inclusion criteria were 1) the study reported an ERP experiment, 2) the study was conducted in human participants, and 3) the study was published in one of the five journals of interest during one of the six years of interest. Exclusion criteria included 1) multimodal studies (e.g., fMRI and ERP), 2) EEG time-frequency analyses (without ERP analyses), 3) poster abstracts, and 4) studies coded as part of the piloting coding procedures for the raters.

Statistical Power Analysis. We had no *a priori* predictions regarding what effect size would be expected for the change in reporting behavior following the publication of the ERP guidelines paper. Alternatively, we conducted sensitivity analyses to determine the statistical effect size that a two-tailed independent samples *t* test would be powered to detect. An independent samples *t* test was chosen in order to compare reporting behavior prior to publication of the guidelines paper to reporting behavior after publication of the guidelines paper. Sensitivity analyses were conducted using a power of .80, a sample size of 75 for each group (pre-guideline publication and post-guideline publication; 150 total articles), and an alpha level of .05. The present analyses had sufficient power to detect a Cohen's *d* of .46 (i.e., a medium effect size).

Article Rating

Rating of the articles followed a rubric that was compiled based on the Keil et al. (2014) guidelines paper. The rubric included information from the following eight categories: Participant Information, EEG Recording, Stimulus and Timing, Preprocessing, ERP measurement, Statistics, Principal Components Analysis (PCA), and Independent Components Analysis (ICA). Information related to sample size, number of groups, statistical analyses, and software packages were also coded.

Each guideline was coded based on whether the information was clearly presented and adequate for confident direct replication of the study. Given the emphasis on estimating observed statistical power of ERP studies, the final sample sizes used for ERP analyses were extracted. The final sample size is typically smaller than the initial sample size reported in ERP studies, because participants are often excluded due to too few trials retained for analysis, poor ERP score reliability, or hardware/software malfunction.

Prior to coding the 150 studies, ten studies were pilot coded by authors PEC and KAC. Mismatches were resolved by unanimous consensus of PEC, KAC, and MJL. The same procedure for coding all 150 articles was followed so that each article was coded by two raters with any discrepancies adjudicated by consensus of all three raters. Percent agreement across variables among the two raters was high (median: 96.7%, mean: 95.0%, minimum to maximum = 79.3% to 100%). Median Cohen's Kappa was acceptable (median: .83, mean: .78, minimum to maximum: .25 to 1.00) but appeared low for a few variables due to the low variability in response options (there were only two rating options for many variables). The pre-registered rubric for article coding, description of specific information relevant to each coded guideline, and the ratings (raw data) are posted on OSF.

Data Analysis

We first investigated whether there were differences in reporting behavior across the five journals using a one-way analysis of variance (ANOVA). Then, we determined whether the ERP guidelines paper impacted reporting behavior. Reporting behavior (i.e., the proportion of reported guidelines) from articles published between 2011and 2013 was compared to reporting behavior from articles published between 2015 and 2017. A two-tailed independent samples *t* test was first conducted to determine whether there was a change in reporting behavior. In order to avoid the biasing effects of heterogeneity of variances, equal variances were not assumed. Hence, Welch's *t* test was used and adjusted degrees of freedom are reported (Welch, 1947). Pooled standard deviations were used in the calculation of Cohen's *d* (Bonett, 2008).

To conclude that there was no meaningful effect of the guidelines paper on reporting behavior, tests of equivalence were performed using the "two one-sided tests" procedure (Lakens, 2017; Schuirmann, 1987). The two one-sided tests procedure provides a framework for estimating that an effect is statistically equivalent to zero or, in other words, that there is "no effect". The equivalence test requires specifying an effect size of interest. Because the current study was only powered to detect a difference for a Cohen's *d* of .46, the test of equivalence used a Cohen's *d* of .50 as the smallest effect size of interest. The two one-sided tests procedure tested whether the difference pre- and post-guidelines reporting behavior was between Cohen's *d*s of -.50 and .50.

Power and sample size estimates for the independent groups *t* test and paired sample *t* test were computed using the `power` command in Stata (version 15.1; StatCorp, 2017). Estimates for the 2-Between x 2-Within group ANOVA were computed using G*Power (version 3; Faul, Erdfelder, Buchner, & Lang, 2009).

Results

Reporting Guidelines

The percentage of guidelines that were reported per article was similar across the five journals, $F(4, 145) = 1.21$, $p = .31$, $\eta^2 = .03$ (see Table 1). Articles reported an average of 63% (*SD* = 7%) of guidelines. The range of reported guidelines across articles was quite wide (range = 39% to 82%), and no article reported all guidelines. As mentioned above, these percentages refer to the proportion of articles that reported the guideline when it was necessary to do so. For example, if an article did not use PCA, it was not necessary to specify any PCA parameters. Considering that reporting behavior was consistent across the five journals (see Table 1), summaries of reporting behavior collapse across journal membership. Information for each journal is separately presented in the supplementary material posted on OSF.

Reporting guidelines were binned into eight categories that were consistent with Keil et al. (2014). Figure 1 shows the percentage of guidelines reported within each category across all 150 articles. In the order of the most guidelines reported to the fewest, Stimulus and Timing (86% of guidelines reported) was the highest reported category and was followed by Participant Information (79%), Statistics (76%), Preprocessing (68%), PCA (65%), EEG Recording (52%), ERP Measurement (51%), and then ICA (36%). The guidelines for each category are discussed in detail below.

Participant information. Participant Information comprises demographic characteristics of the participants, which includes gender, age, and education level of participants (see Figure 2). Most articles reported the gender (99%) and age (95%) of participants, but fewer articles reported education level (43%).

EEG recording. The EEG Recording category consisted of guidelines related to the online recording of EEG, such as information about the sensors, amplifiers, and online filtering

(see Figure 2). Most articles reported the EEG sampling rate (97%) and the online reference electrodes (87%). Although the majority of articles also reported the online filter cutoffs (75%), very few articles reported specific information about filter characteristics. Most articles failed to report whether half-amplitude or half-power cutoffs were used (4%), the online filter roll-off (3%), or filter family (2%).

Stimulus and timing. This category refers to the stimulus and timing parameters of the paradigm used during EEG recording (see Figure 2). Most articles reported clear information about the timing characteristics of the paradigm (90%). However, only 56% of articles reported enough information about the stimuli (such as specific information about color, size, or which pictures were used) that would allow direct replication of the paradigm to be possible.

Preprocessing. The Preprocessing category comprises information related to offline EEG data reduction and processing (see Figure 2). The order in which preprocessing steps was performed was clear in 90% of articles. When applicable, all articles (100%) clearly reported the offline reference and the offline filter cutoffs. However, few articles reported whether half-amplitude or half-power cutoffs (2%) were used or the filter roll-off (26%) and filter family (18%).

Preprocessing steps are implemented in software packages, and the reporting of such information can sometimes be used to infer some specific parameters of preprocessing. Of the 150 articles coded, 86 (57%) reported the software packages used for EEG data analysis, and some of these articles reported using multiple software packages (see Table 2). There were 17 different software packages reported, and the most frequently used software packages were EEGLab (n = 31) and BrainVision Analyzer (n = 28).

ERP measurement. This category mostly consisted of information related to ERP quantification (see Figure 3). Most articles reported inferential statistics (97%), measurement sensors (95%), measurement timing window (95%), and the measurement approach (91%). However, very few articles reported whether *a priori* sensors (5%) and temporal windows (3%) were used. For peak amplitude measures, only 8% of articles reported whether a local or absolute peak amplitude approach was used.

Statistical analyses. This category refers to information related to statistical analyses (see Figure 3). Most articles reported *p* values (99%), clearly described the statistical procedures used (95%), and provided inferential test statistics (91%). When permutation statistics were used, about half of articles reported the number of permutations (44%) and the method for identifying a significance threshold (56%). Appropriate corrections for violating assumptions of statistical models was reported in 55% of articles, and 40% of articles considered corrections for multiple comparisons.

PCA. When PCA was used, all articles (100%) provided sufficient information so that the preprocessing steps implemented prior to PCA were clear (see Figure 3). Most articles described the structure of EEG data submitted to the PCA (75%), the rotation applied to the data (62%), and the decision rule for retaining or discarding components (62%). Half of the articles (50%) described the PCA algorithm and relatively few articles described the type of association matrix (38%).

ICA. When ICA was used, all articles provided sufficient information regarding preprocessing steps prior to the ICA (see Figure 3). The majority of articles described how ICA components were selected (58%). However, very few articles reported the ICA algorithm (13%),

structure of the data submitted to the ICA (8%), or the number of components retained/removed (3%).

Impact of Guidelines Paper

To determine whether the publication of guidelines for ERP studies impacted reporting behavior, reporting behavior for the three years prior to the publication of the guidelines paper was compared to the reporting behavior for the three years following its publication (see Table 3). There was no significant difference in reporting behavior, $t(142.33) = 0.73$, $p = .47$, Cohen's $d = .12$, 95% CI [-1.5%, 3.2%].

Beyond determining whether the guidelines paper increased or decreased reporting behavior, analyses were performed to determine if the results were statistically equivalent to the absence of an effect. The equivalence tests set equivalence bounds to $\pm$ 3.62%. Both lower-bound and upper-bound equivalence tests were significant, $t(142.33) = 2.33$, $p = .01$; $t(142.33) = -3.79$, $p < .01$, 95% CI [-2.8%, 1.1%], respectively.

Based on the combination of the null-hypothesis significant test and the equivalence tests, the observed impact of the guidelines paper was not statistically different from zero and was statistically equivalent to zero. In short, the publication of the ERP guidelines paper had no impact on reporting behavior in the three years after publication.

Sample Sizes

Summary statistics for the number of participants examined in each article are presented in Table 1. For all articles coded, each article contained an average of 29 (median = 22, $SD = 21$, range 5 to 146) participants. The sample size decreased when considering the number of participants *in each group* examined in each article. When considering participants per group, each article contained an average of 21 (median = 18, $SD = 11$; range = 5 to 86) participants per

group. Of those articles that reported data on more than one group, 77% examined two groups of participants.

Statistical Power

There was a great deal of heterogeneity in the statistical models used in the coded articles. A summary of the between- and/or within-subject ANOVA models used in at least three articles are shown in Table 4 (see supplementary material on OSF for a description of all statistical models). In order to estimate the statistical power in the coded articles, we estimated power for three models: independent samples *t* tests, paired samples *t* tests, and a 2-Between Group x 2-Within Group ANOVA interaction (see Figure 4 and Table 5). Additional power analyses were not conducted for the 2-Within Group x 2-Within Group ANOVA design, because it is equivalent to a paired *t* test on difference scores.

The number of participants needed to achieve a statistical power of .80 for a large effect size (Cohen's $d = .8$, Cohen's $f = .4$) and an alpha level of .05 was 52 participants (26 participants per group) for an independent samples *t* tests, 15 participants for a paired samples *t* test, and 22 participants (11 participants per group) for a 2-Between Group x 2-Within Group ANOVA interaction. The number of participants needed to achieve a statistical power of .80 for a small effect size (Cohen's $d = .2$, Cohen's $f = .1$) and an alpha level of .05 was 788 participants (394 participants per group) for an independent samples *t* test, 199 participants for a paired samples *t* test, and 298 participants (194 participants per group) for a 2-Between Group x 2-Within Group ANOVA interaction. A summary of the number of participants needed for each effect size and statistical design is provided in Table 6.

Next, the average sample size in the coded articles was considered to determine the average effect size coded articles were powered to detect. The effect size (Cohen's *d* or Cohen's

f) for a two-tailed test that a study with 21 participants per group, a statistical power of .80, and an alpha level of .05 would be able to detect is .89 (large effect size) for an independent samples *t* tests, .62 (medium-to-large effect size) for a paired samples *t* test, and .44 (large effect size) for 2-Between Group x 2-Within Group ANOVA interaction.

Lastly, the achieved statistical power was computed based on a sample size of 21 participants per group (see Table 7), and this analysis provides a conservative estimate of the statistical power of coded ERP studies. Studies only achieved a power of .72 to detect a large effect size (Cohen's *d* =.80) for an independent samples *t* test. Studies that used paired samples *t* tests achieved a power of .94 to detect a large effect size (Cohen's *d* = .80) but were insufficiently powered (achieved power: .59) to detect a medium effect size (Cohen's *d* = .50). For a 2-Between Group x 2-Within Group interaction, studies achieved .98 power to detect a large effect size (Cohen's *f* = .40), but they were underpowered (achieved power: .73) to detect a medium effect size (Cohen's *f* = .25). Taken together, independent samples *t* tests were insufficiently powered to detect large effect sizes, and paired samples *t* tests and tests of the 2-Between Group x 2-Within Group interactions were only sufficiently powered to detect large effect sizes.

Discussion

Across 150 ERP studies, an average of 63% of guidelines were reported, which suggests that published ERP studies omit key information required for independent replication. This reporting behavior was consistent across five prominent ERP journals: *Clinical Neurophysiology*, *International Journal of Psychophysiology*, *Journal of Cognitive Neuroscience*, *NeuroImage*, and *Psychophysiology*. Hence, gaps in methods reporting appear to be a shortcoming of the field, rather than any specific journal or impact factor level. Notably, the

ERP guidelines paper (Keil et al., 2014) had no impact on reporting behavior for the three years following its publication. With regard to the sample size of ERP studies, the average sample size per group was 21 participants. Considering this sample size, ERP studies had sufficient power to observe only large statistical effects (Cohen's $d > .8$, Cohen's $f > .4$). Taken together, the present study revealed critical shortcomings in the reporting of common ERP practices that hinder the ability to independently replicate ERP studies and the probability for ERP studies to find replicable effects.

Reporting Behavior

The widespread omission of over a third of the required reporting guidelines serves as a substantial barrier to replication efforts and to the evaluation of research quality. It is unclear how to judge whether experimental manipulations and data collection and processing are sound without sufficient details to be reproducible. Poor research reporting documentation is not unique to ERP studies and has been observed in other subfields of neuroscience (Carp, 2012; Guo et al., 2014; Muncy, Hedges-Muncy, & Kirwan, 2017; Poldrack et al., 2017) and the biomedical sciences (Chalmers & Glasziou, 2009; Glasziou et al., 2014; Ioannidis et al., 2014). It is estimated that in the biomedical science, billions of dollars are wasted every year due to the consequences of misreporting and inadequate reporting of research (Chalmers & Glasziou, 2009). It is likely that missing details in the ERP data analysis pipeline might similarly lead to wasted resources.

A potential reason that Methods sections lack enough information for replication is an underappreciation of the importance of direct (close) replications (Chambers, 2017; Nosek & Lakens, 2014; Schmidt, 2009; Simons, 2014). A direct replication tests the repeatability of a finding by duplicating the study design and analysis of the original study. Direct replications are

important for increasing the precision of effect size estimates in meta-analyses, establishing generalizability of effects, identifying boundary conditions for "real" effects, and correcting scientific theory (Nosek & Lakens, 2014). It is important for direct replications to be conducted by outside laboratories in order to verify the robustness of the effects found in the original research (Nosek & Lakens, 2014; Schmidt, 2009; Simons, 2014). Without reporting all experimental details, it is unlikely that replication studies will be successful. In fact, failure to report important methodological details that impact study findings or failure to disclose flexibility in data analysis is considered a "questionable" research practice that contributes to replication difficulties (Forstmeier et al., 2017; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011).

Unfortunately, direct replications are quite rare, and in psychology they are frequently replaced with conceptual replications (LeBel & Peters, 2011; Nosek & Lakens, 2014; Schmidt, 2009). A conceptual "replication"[1] seeks to test a phenomenon using a different method, and conceptual replications are essential for theory testing. The consequence of replacing direct replications with conceptual replications is that conceptual replications cannot conclusively disprove the original finding and failures to replicate are often attributed to methodological changes (LeBel & Peters, 2011; Simons, 2014). Hence, science cannot self-correct when theories are built on conceptual replications. Additionally, conceptual replications can exploit researcher degrees of freedom to erroneously support the original study idea by changing analysis approaches (Forstmeier et al., 2017; Simmons et al., 2011; Simons, 2014).

Some information was consistently underreported across coded studies, suggesting an underappreciation of the impact of reporting some EEG data reduction parameters on final ERP scores. The most frequently underreported parameters related to characteristics of the online and

offline filters. The ERP guidelines paper recommends reporting the filter cutoffs, specifying whether those cutoffs were half-amplitude or half-power, describing the filter roll-offs, and identifying the filter family. The most commonly reported aspect of filtering was the filter cutoff. However, even when using the same filter cutoffs, the signal quality is differentially impacted based on the other mentioned filter characteristics (Widmann & Schröger, 2012; Widmann, Schröger, & Maess, 2015). The characteristics of the filter used should be determined by the ERP components of interest, and the same filter cutoffs and filter characteristics are not well suited to all studies of any ERP component (Cook & Miller, 1992; Edgar, Stewart, & Miller, 2005; Nitschke, Miller, & Cook, 1998; Widmann & Schröger, 2012; Widmann et al., 2015).

It was also rare for a study to report using *a priori* determined sensors and temporal windows for ERP measurement. A common approach to scoring ERPs is to choose sensors and temporal windows based on grand averages where the effect of interest appears maximal. This practice often leads to finding significant effects but results in a high rate of spurious findings (Luck & Gaspelin, 2017). When possible, it is considered "best practice" to select sensors and temporal windows *a priori* to avoid biased measurement and analysis. However, in some cases it is not possible to have definite *a priori* predictions, such as when using a novel paradigm. In these instances, there are alternative approaches, such as using a functional localizer, collapsed localizer, a window-independent or mass univariate measurement approach, or factor analysis (see Dien, 2017; Luck & Gaspelin, 2017). Regardless of the approach used, the measurement sensors and temporal windows used should be clearly reported and justified (Keil et al., 2014).

Sample Size and Statistical Power

The average overall sample size of studies included was 29 (median = 22); the typical sample size of a group of participants was 21 on average (median = 18) in the coded articles, and

this resulted in an estimated statistical power of .72-.98 for a large effect size, .35-.73 for a medium effect size, and .10-.18 for a small effect size (see Table 7). However, these power estimates are considered conservative, because the majority of studies used more complicated statistical designs that will reduce power. Regardless, the observed sample size and statistical power for the statistical designs of focus are similar to other subfields of neuroscience. For example, estimates of the median sample sizes of fMRI studies range from 15 (Carp, 2012) to 28.5 (Poldrack et al., 2017) for single-group studies and from 14.75 (Carp, 2012) to 19 (Poldrack et al., 2017) per group for studies with multiple groups. The estimated statistical power of fMRI studies is between .08 and .31 (Button et al., 2013b), and a recent large assessment of statistical power in the fields of cognitive neuroscience and psychology estimated a median power of .73 for large effects, .44 for medium effects, and .12 for small effects (Szucs & Ioannidis, 2017). Although statistical power of ERP studies is on par with the fields of cognitive neuroscience and psychology, the consequences of small samples and low power nonetheless limit the interpretability and potential replicability of ERP studies.

The coded ERP studies were only powered at a level of .80 to detect large statistical effects for paired samples *t* tests and 2-Between Group x 2-Within Group ANOVA interactions, which together accounted for only 15% of coded studies. Given that statistical power was low for studies of smaller effects and that most studies used more complicated statistical analyses, many observed statistical effect sizes are likely exaggerated due to the statistical significance threshold commonly applied to published studies (Gelman, 2018; Rosenthal, 1979; Simmons et al., 2011). This bias to publish statistically significant effects incentivizes studying small samples and noisy measurements, because researchers can exploit the garden of forking paths (or researcher degrees of freedom) to find statistically significant effects (Baldwin, 2017; Brand & Bradley, 2016;

Clayson & Miller, 2017b; Gelman, 2018; Gelman & Carlin, 2014; Larson & Carbine, 2017; Loken & Gelman, 2017).

When researcher degrees of freedom are intentionally exploited, such as when multiple iterations in the data processing pipeline are tested until a significant result is obtained, the likelihood of finding replicable effects is reduced. Furthermore, some meta-analytic approaches for estimating effect sizes are unable to adjust for the presence of this type bias. Most of these approaches are only designed to adjust for journal publication bias (i.e., the bias that journals are more likely to publish significant studies than non-significant studies), but they are not designed to adjust for questionable research practices that inflate the likelihood of finding statistical significance (Carter, Schönbrodt, Gervais, & Hilgard, 2018; Simonsohn, Nelson, & Simmons, 2014a, 2014b; Simonsohn, Simmons, & Nelson, 2015). As a result, some meta-analytic approaches are unable to identify the "true" effect size in the literature and suffer from inflated false positives (Carter et al., 2018). Hence, ERP meta-analyses might consider employing approaches that identify whether the literature is biased due to questionable research practices, such as undisclosed researcher degrees of freedom. A *p*-curve analysis is one such approach that operates under the assumption of journal publication bias and can identify the use of questionable research practices to obtain statistically significant effects (Clayson, Carbine, & Larson, in principle acceptance; Simonsohn et al., 2015). Indeed, a pre-registered *p*-curve analysis on current and 10-year-past psychophysiological studies showed generally good evidential value and low selective reporting in the field, but demonstrated relatively low average statistical power (Carbine, Lindsey, Rodeback, & Larson, 2019).

The estimate of statistical power of the coded articles is considered conservative, because it was based on the most frequently used approaches for statistical analysis. For example, most

studies used more factors or levels in ANOVAs than were considered in the power analyses, which would lead to lower power (see Table 4) and do not appear to follow best practices for repeated measures ANOVAs of ERP data (Dien, 2017). Including many factors in exploratory ANOVAs also leads to an increase in the family wise error rate, because it is not common practice to correct for multiple comparisons in ANOVAs (Cramer et al., 2015; Luck & Gaspelin, 2017). Luck and Gaspelin (2017) showed that the probability of a Type I error is 5%, 14%, 30%, 54%, and 80% for a one-, two-, three-, four-, and five-factor ANOVA, respectively. For the coded articles, most of the articles used more than one factor (see Table 4), indicating that the familywise error rate is above 5%. ERP studies[2] often conduct many ANOVAs, such as an ANOVA on multiple ERP components, on amplitude and latency measurements, or across multiple time windows or electrodes. For example, some coded ERP studies performed multifactorial ANOVAs on 50ms chunks of activity across the entire ERP epoch (e.g., -200 to 800ms, resulting in 20 separate ANOVAs). Such practices virtually ensure that a statistically significant, although possibly spurious or inflated, effect will be observed in an ERP study. When conducting ANOVAs, correcting for multiple comparisons, reducing the number of factors, and removing unnecessary analyses should be used to reduce the familywise and experiment-wise error rates (see Luck & Gaspelin, 2017).

Limitations

The present study has some limitations. We only coded whether authors followed guidelines based on what was stated by the authors in the published studies. It is possible that additional analysis steps were performed, but not reported. Some guidelines, such as the type of interpolation used or offline filtering, were only coded when they were reported, but it is likely that additional unreported steps were performed by authors. For example, some studies reported

conducting topographical analyses but did not mention how bad channels were interpolated or handled. In such instances, the type of interpolation was not coded, even though it is likely that interpolation was conducted. Hence, it is likely that the reporting behavior presented here was somewhat overestimated. The reported analyses focused on ERP studies, but other types of EEG studies, such as time-frequency analyses, might more or less closely adhere to reporting guidelines (Cohen, 2017). It is also possible that the publication of the Keil et al. (2014) guidelines paper had no impact on reporting behavior, because reporting behavior has remained stable since the publication earlier ERP reporting guidelines (Donchin, Callaway, Cooper, & Desmedt, 1977; Picton et al., 2000; Pivik et al., 1993). Nonetheless, only two thirds of the required information for replicating contemporary ERP studies is being routinely reported.

Furthermore, the spirit of the Keil et al. (2014) guidelines paper was to facilitate communication among researchers by providing explicit recommendations for reporting (see How To Use This Document section; Keil et al., 2014, p. 2). It is possible that authors consciously chose to deviate from the recommended guidelines for reasons specific to their study, which is acceptable. However, "such deviations are [to be] explicitly documented and explained" (Keil et al., 2014, p. 2). In the present coding procedure, a guideline was coded as acceptable as long as it was addressed. For example, sensors did not need to be chosen *a priori* so long as how sensors would be chosen was specified, such as through a mass univariate or functional localizer approach. Hence, it is unlikely that such deviations accounted for the low reporting behavior. In line with the spirit of the guidelines paper, we believe that communication of methodological parameters among researchers could, and should, be improved.

For the present manuscript we randomly selected articles from five different journals; these articles examined healthy participants as well as various clinical and developmental

populations, but this was not explicitly coded. It is possible that some populations require special considerations when recording and analyzing ERP data. Consistent with recommendations from the guidelines paper, deviations from standard protocol should be documented and justified. Although such data can be costly and difficult to acquire, "the rules of statistical inference have no empathy for how hard it is to acquire data" (Nosek, Ebersole, DeHaven, & Mellor, 2018, p. 5). Despite that data collection might be slow, the driving research questions are important enough to answer rigorously.

Moving Forward

Failing to report all key methodological parameters appears to be commonplace in ERP research, which serves as a substantial barrier to replication efforts. Because each methodological parameter can impact ERP findings, reporting all parameters is a "best practice" for evaluating research quality and replicability. The question remains as to how does the field move forward to resolve these issues. We offer a few suggestions below based on how other fields are addressing the replication problem in science (see Button et al., 2013a, 2013b; Ioannidis, 2005; Ioannidis, 2008; Ioannidis et al., 2014; Lilienfeld, 2017; Tackett, Brandes, King, & Markon, 2019; Yom, 2018) and discuss some issues that are specific ERP research.

The disclosure of key methodological parameters will increase transparency that hopefully unveils when researcher degrees of freedom are exploited and calibrates a careful reader's confidence in reported effects. As such, it would be helpful for editors and reviewers to enforce ERP reporting guidelines. A reason that key information might be omitted from ERP studies is due to space limitations for journal articles. In such an event, authors could be encouraged to post all study details necessary for direct replication as supplementary material or to online repositories, such as OSF. Open sharing of processing code and experimental tasks

could also enhance reporting of most pipeline steps and further facilitate replication. The ERP guidelines paper has a checklist in the Appendix for ensuring that all key parameters are reported (Keil et al., 2014). A completed checklist could be submitted with journal articles to ensure that all methodological parameters are communicated in the manuscript.

It is possible that some researchers are carefully considering each methodological parameter and simply not reporting each parameter due to oversight. Alternatively, researchers might be exploiting researcher degrees of freedom to find statistically significant effects. One approach that combats the exploitation of researcher degrees of freedom is study preregistration or the registered reports format adopted in some journals (Larson, 2016; Munafò et al., 2017; Nosek et al., 2018; Nosek & Lakens, 2014). In essence, a preregistration is a locked analysis plan that is sealed before any data analysis (and ideally data collection) is conducted. Preregistration, when correctly followed, prevents the exploitation of researcher degrees of freedom by locking in a data analysis plan prior to examining the data (Nosek et al., 2018). Preregistrations of ERP studies can include a prespecified hypothesis, EEG preprocessing plan, and a data analysis plan, and it could be helpful to complete the checklist from the Appendix of the ERP guidelines paper in such preregistrations. The OSF offers a mechanism for preregistering a study (https://osf.io/) that can include the information mentioned above as well as software scripts for data processing and statistical analysis.

The gold standard of preregistration is the registered report format (https://cos.io/rr/), which consists of two phases of peer review. In the first phase, the study hypothesis and methodology are peer reviewed prior to data collection. Upon successful completion of this first phase, the manuscript is provisionally accepted for publication. During the second phase, the full manuscript is reviewed and published as long as the proposed methodology was followed. This

format removes one incentive (the publication of a manuscript) to exploit researcher degrees of freedom at the data analysis stage, because the manuscript is accepted for publication at the completion of the first phase regardless of whether significant effects are observed. Of the primary psychophysiology journals, only the *International Journal of Psychophysiology* has, thus far, implemented the registered reports format (Larson, 2016). Many have argued that the incentive structure for academia needs to shift away from rewarding voluminous publishing to rewarding rigorous, careful research (Baldwin, 2017; Ioannidis et al., 2014; Nelson, Simmons, & Simonsohn, 2012). An advantage of a registered report format is that it shifts the incentive away from massaging data to uncover a statistically significant effect to designing a careful test of a specific hypothesis.

Another barrier to reporting might be that researchers do not know the specific ways in which data were processed and analyzed due to a reliance on various software analysis packages (for a similar discussion, see Software as a Black Box section in Clayson & Miller, 2017b). There are numerous software packages available for processing and analyzing ERP data, and it is impossible to be an expert in all methodologies. Hence, the appeal of prepackaged code is understandable, but such code can become a 'black box' (Clayson & Miller, 2017b). The extent to which popular software analysis packages can build in validity checks would be helpful for researchers who are not easily able to judge when analysis approaches are appropriate. Another useful feature for popular software packages that process ERP data would be functions that generate a print out of how the data were processed. Ideally such a processing summary would mirror the Appendix of the ERP guidelines paper (Keil et al., 2014) and provide all information that should be reported in a Methods section.

Although it is easy to suggest that researchers collect more participants to improve statistical power, there may be practical barriers that prevent some from doing so (e.g., limited access to particular patient populations). One approach to improve statistical power is through collaboration by conducting multisite ERP studies (Baldwin, 2017), and this is already a popular approach among some fMRI groups. Such multisite studies can increase samples sizes, statistical power, and generalizability of findings. In addition to multisite studies, depositing EEG data in repositories can facilitate sharing and combining of datasets to hopefully improve statistical power. A few repositories currently exist for such purpose and include the OpenfMRI database (https://www.openfmri.org/; Poldrack et al., 2013; Poldrack & Gorgolewski, 2017) and the Patient Repository for EEG Data + Computational Tools (PRED+CT; http://predict.cs.unm.edu/; Cavanagh, Napolitano, Wu, & Mueen, 2017). Furthermore, funding agencies have started to require depositing data to repositories. For example, all grant applications and awards submitted to the National Institute of Mental Health (NIMH) that involve human subjects after January 1, 2020 will be required to deposit all raw and analyzed data, including psychophysiological data, unless an explicit exception is granted (https://grants.nih.gov/grants/guide/notice-files/NOT-MH-19-033.html).

Along the lines of statistical power, one feature that is often underappreciated when conducting *a priori* power calculations is the impact of score reliability on statistical power. Unreliable scores can reduce statistical power (Boudewyn, Luck, Farrens, & Kappenman, 2017; Clayson & Miller, 2017b; Fischer, Klein, & Ullsperger, 2017; Kolossa & Kopp, 2018; Luck & Gaspelin, 2017). Given the relationship between the number of trials included in an ERP average and internal consistency estimates of reliability (Clayson & Miller, 2017a), power contour plots can be used to estimate the optimal balance between the number of experimental trials, the

number of participants, and the statistical power of a given effect size (Baker et al., 2019). The positive relationships between reliability and effect sizes have been shown in both between-group (Hajcak, Meyer, & Kotov, 2017) and within-person (Clayson & Miller, 2017a) ERP studies. For example, between-group effect sizes (healthy controls vs. people with generalized anxiety disorders) increased with increases in internal consistency (Hajcak et al., 2017).

Conclusions

An average of 63% of key methodological parameters were reported across 150 ERP studies from five prominent journals, which suggests that the underreporting of recommended guidelines is a ubiquitous practice. Hopefully, this underreporting is due to oversight on the part of authors. However, it is possible that underreporting might be due to attempts to obscure data processing and analysis practices when authors exploit researcher degrees of freedom to find statistically significant effects. We have recommended some solutions, such as the preregistration of data processing and analysis plans to motivate rigor over novelty. We hope that moving forward authors, reviewers, and editors encourage the use of the ERP reporting guidelines (Keil et al., 2014) to facilitate communication among researchers and improve the replicability of ERP research.

Small sample sizes and low statistical power appear endemic to ERP studies, which is consistent with the larger field of cognitive neuroscience. Our findings suggest that ERP research is powered to detect only large statistical effects for simple statistical designs. Anecdotally speaking, one of the advantages of using ERPs over some other neuroimaging techniques, such as functional magnetic resonance imaging (fMRI), is that conducting ERP studies is more affordable. The affordability is often cited as an advantage, because larger sample sizes can be obtained for a lower cost. Although ERP research is indeed more affordable, the typical ERP

study seems to suffer from the same problems associated with small sample sizes as fMRI studies, and this fact appears to be underappreciated, at least anecdotally. Small samples, low statistical power, and undisclosed researcher flexibility contribute the low replicability ERP studies. The replicability of ERP studies can be improved by conducting *a priori* sample size calculations to ensure adequately powered samples for relevant effect sizes and by conducting multisite ERP studies to increase sample size and generalizability of findings.

References

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2019). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *ArXiv*.

Baldwin, S. A. (2017). Improving the rigor of psychophysiology research. *Int J Psychophysiol, 111*, 5-16. doi:10.1016/j.ijpsycho.2016.04.006

Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13*, 99-109. doi:10.1037/1082-989X.13.2.99

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2017). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology, 14*, e13049. doi:10.1111/psyp.13049

Brand, A., & Bradley, M. T. (2016). The precision of effect size estimation from published psychological research: Surveying confidence intervals. *Psychological Reports, 118*, 154-170. doi:10.1177/0033294115625265

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013a). Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience, 14*, 585-586. doi:10.1038/nrn3475-c4

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

Carbine, K. A., Lindsey, H. M., Rodeback, R. E., & Larson, M. J. (2019). Quantifying evidential value and selective reporting in recent and 10-year past psychophysiological literature: A pre-registered P-curve analysis. *International Journal of Psychophysiology, 142*, 33-49. doi:10.1016/j.ijpsycho.2019.06.004

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage, 63*, 289-300. doi:10.1016/j.neuroimage.2012.07.004

Carp, J. (2013). Better living through transparency: Improving the reproducibility of fMRI results through comprehensive methods reporting. *Cognitive, Affective, & Behavioral Neuroscience, 13*, 660-666. doi:10.3758/s13415-013-0188-0

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2018). Correcting for bias in psychology: A comparison of meta-analytic methods. *PsyArXiv*. doi:10.31234/osf.io/9h3nu

Cavanagh, J. F., Napolitano, A., Wu, C., & Mueen, A. (2017). The patient repository for EEG data + computational tools (PRED+CT). *Frontiers in Neuroinformatics, 11*, 67. doi:10.3389/fninf.2017.00067

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *Lancet, 374*, 86-89. doi:10.1016/S0140-6736(09)60329-9

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*: Princeton University Press.

Clayson, P. E., Carbine, K. A., & Larson, M. J. (in principle acceptance). Error-related negativity and reward positivity as biomarkers of depression: P-curving the evidence. *International Journal of Psychophysiology*.

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

Clayson, P. E., & Miller, G. A. (2017b). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Psychophysiology, 111*, 57-67. doi:10.1016/j.ijpsycho.2016.09.005

Cohen, M. X. (2017). Rigor and replication in time-frequency analyses of cognitive electrophysiology data. *International Journal of Psychophysiology, 111*, 80-87. doi:10.1016/j.ijpsycho.2016.02.001

Cook, E. W., III, & Miller, G. A. (1992). Digital filtering: Background and tutorial for psychophysiologists. *Psychophysiology, 29*, 350-362. doi:10.1111/j.1469-8986.1992.tb01709.x

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., . . . Wagenmakers, E.-J. (2015). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review, 23*, 640-647. doi:10.3758/s13423-015-0913-5

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin, 144*, 757-777. doi:10.1037/bul0000154

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. doi:10.1016/j.ijpsycho.2016.09.006

Donchin, E., Callaway, E., Cooper, E., & Desmedt, R. (1977). Publication criteria for studies of evoked potentials (EP) in man. Report of the methodology committee. In J. E. Desmedt (Ed.), *Progress in clinical neurophysiology: Vol 1. Attention, voluntary contraction and event-related cerebral potentials* (Vol. 1, pp. 1-11). Basel, Switzerland: Karger.

Edgar, J. C., Stewart, J. L., & Miller, G. A. (2005). Digital filters in ERP research. In T. C. Handy (Ed.), *Event-Related Potentials: A Methods Handbook* (pp. 33-56). Cambridge, Massachusetts: MIT Press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. doi:10.3758/BRM.41.4.1149

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. doi:10.1111/psyp.12863

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings - a practical guide. *Biological Reviews of the Cambridge Philosophical Society, 92*, 1941-1968. doi:10.1111/brv.12315

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin, 44*, 16-23. doi:10.1177/0146167217729162

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641-651. doi:10.1177/1745691614551642

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., . . . Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *Lancet, 383*, 267-276. doi:10.1016/S0140-6736(13)62228-X

Guo, Q., Parlar, M., Truong, W., Hall, G., Thabane, L., McKinnon, M., . . . Pullenayegum, E. (2014). The reportinrg of observational clinical functional magnetic resonance imaging

studies: A systematic review. *PloS one, 9*, e94412-94411. doi:10.1371/journal.pone.0094412

Hajcak, G., Meyer, A., & Kotov, R. (2017). Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. *Journal of Abnormal Psychology, 126*, 823-834. doi:10.1037/abn0000274

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

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640-648. doi:10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., . . . Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *Lancet, 383*, 166-175. doi:10.1016/S0140-6736(13)62227-8

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532. doi:10.1177/0956797611430953

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

Kolossa, A., & Kopp, B. (2018). Data quality over data quantity in computational cognitive neuroscience. *NeuroImage, 172*, 775-785. doi:10.1016/j.neuroimage.2018.01.005

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*, 355-362. doi:10.1177/1948550617697177

Larson, M. J. (2016). Commitment to cutting-edge research with rigor and replication in psychophysiological science by Michael J. Larson. *International Journal of Psychophysiology, 102*, ix-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. doi:10.1016/j.ijpsycho.2016.06.015

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*, 371-379. doi:10.1037/a0025172

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science, 12*, 660-664. doi:10.1177/1745691616687745

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*, 584-585. doi:10.1126/science.aal3618

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. doi:10.1111/psyp.12639

Moser, J. S., Moran, T. P., Kneip, C., Schroder, H. S., & Larson, M. J. (2016). Sex moderates the association between symptoms of anxiety, but not obsessive compulsive disorder, and error-monitoring brain activity: A meta-analytic review. *Psychophysiology, 53*, 21-29. doi:10.1111/psyp.12509

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behavior, 1*, 1-9. doi:10.1038/s41562-016-0021

Muncy, N. M., Hedges-Muncy, A. M., & Kirwan, C. B. (2017). Discrete pre-processing step effects in registration-based pipelines, a preliminary volumetric study on T1-weighted images. *PloS one, 12*, e0186071-0186015. doi:10.1371/journal.pone.0186071

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology, 69*, 511-534. doi:10.1146/annurev-psych-122216-011836

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2012). Let's publish fewer papers. *Psychological Inquiry, 23*, 291-293. doi:10.1080/1047840X.2012.705245

Nitschke, J. B., Miller, G. A., & Cook, E. W., III. (1998). Digital filtering in EEG/ERP analysis: Some technical and methodological comparisons. In M. Eimer (Ed.), *Behavior Research Methods, Instruments, and Computers* (Vol. 30, pp. 54-67).

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The Preregistration Revolution. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 2600-2606. doi:10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137-141. doi:10.1027/1864-9335/a000192

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

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.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., . . . Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*, 115-126. doi:10.1038/nrn.2016.167

Poldrack, R. A., Barch, D. M., Mitchell, J. P., Wager, T. D., Wagner, A. D., Devlin, J. T., . . . Milham, M. P. (2013). Towards open sharing of task-based fMRI data: The OpenfMRI project. *Frontiers in Neuroinformatics, 7*, 12. doi:10.3389/fninf.2013.00012

Poldrack, R. A., & Gorgolewski, K. J. (2017). OpenfMRI: Open sharing of task fMRI data. *NeuroImage, 144*, 259-261. doi:10.1016/j.neuroimage.2015.05.073

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Schmidt, S. J. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90-100. doi:10.1037/a0015108

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics, 15*, 657-680.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*, 487-510. doi:10.1146/annurev-psych-122216-011845

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366. doi:10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*, 76-80. doi:10.1177/1745691613514755

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). p-Curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666-681. doi:10.1177/1745691614553988

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). p-Curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534-547. doi:10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144*, 1146-1152. doi:10.1037/xge0000104

StatCorp. (2017). Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*, e2000797-2000718. doi:10.1371/journal.pbio.2000797

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology, 15*, 579-604. doi:10.1146/annurev-clinpsy-050718-095710

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika, 34*. doi:10.2307/2332510

Widmann, A., & Schröger, E. (2012). Filter effects and filter artifacts in the analysis of electrophysiological data. *Frontiers in Psychology, 3*, 233. doi:10.3389/fpsyg.2012.00233

Widmann, A., Schröger, E., & Maess, B. (2015). Digital filter design for electrophysiological data--A practical approach. *Journal of Neuroscience Methods, 250*, 34-46. doi:10.1016/j.jneumeth.2014.08.002

Yom, S. (2018). Analytic transparency, radical honesty, and strategic incentives. *Political Science & Politics, 51*, 416-421. doi:10.1017/S1049096517002554

Footnote

1. Using the term replication when referring to a study of the robustness of an effect under different methodological parameters is a misnomer. Nothing is being replicated per se. Rather a phenomenon of interest is simply being tested under different conditions.
2. We chose not to provide specific citations as examples in the Discussion section. All of the issues discussed occurred in multiple articles, and thus it is likely a field-wide issue rather than an issue with one particular group or lab.

Acknowledgements

Writing of this manuscript was supported in part by the Office of Academic Affiliations, Advanced Fellowship Program in Mental Illness Research and Treatment, Department of Veterans Affairs. Michael J. Larson, PhD, is the Editor-in-Chief of the *International Journal of Psychophysiology* and receives an honorarium from Elsevier, the journal publisher. *International Journal of Psychophysiology* was one of the journals reviewed for the present manuscript. Otherwise, we have no other acknowledgements, conflicts of interest, or grant funding sources to report.

Table 1

Summary Statistics for Reported Guidelines and Sample Sizes

	Percent of Reported Guidelines Per Article			
Journal	Mean	Median	*SD*	Range
Clinical Neurophysiology	63%	65%	10%	39% to 82%
International Journal of Psychophysiology	64%	65%	7%	51% to 77%
Journal of Cognitive Neuroscience	62%	62%	6%	51% to 72%
NeuroImage	61%	61%	7%	44% to 72%
Psychophysiology	64%	65%	6%	53% to 74%
All Journals	63%	64%	7%	39% to 82%

	Total Participants Per Article			
Journal	Mean	Median	*SD*	Range
Clinical Neurophysiology	43	33	*32*	5 to 146
International Journal of Psychophysiology	27	22	17	8 to 80
Journal of Cognitive Neuroscience	23	21	12	10 to 66
NeuroImage	20	17	8	10 to 46
Psychophysiology	31	24	22	10 to 96
All Journals	29	22	21	5 to 146

	Total Participants Per Group Per Article			
Journal	Mean	Median	*SD*	Range
Clinical Neurophysiology	25	21	17	5 to 86
International Journal of Psychophysiology	18	17	7	8 to 55
Journal of Cognitive Neuroscience	20	17	10	9 to 66
NeuroImage	18	16	5	10 to 27
Psychophysiology	21	21	7	10 to 40
All Journals	21	18	11	5 to 86

Note: Total participants per article indicates the number of participants in each coded article (ignoring group membership). Total participants per group per article indicates the number of participants in each group for a given article. The 'All Journals' row represents the summary statistics across all journals.

Table 2

Frequency Table of Software Packages

Software Package	Frequency
EEGLab	31
BrainVision Analyzer	28
Brain Electrical Source Analysis (BESA)	7
Cartool	6
Scan	6
ERP PCA (EP) Toolkit	5
Fieldtrip	5
ERPLab	4
EEProbe	3
NetStation	2
BrainStorm	1
EMSE	1
EPlyzer	1
ERPSS	1
Fully Automated Statistical Thresholding for EEG Artifact Rejection (FASTER)	1
Statistical Parametric Mapping (SPM)	1

Table 3

Percentage of Guidelines Reported by Year

Year	Mean	Median	*SD*	Range
2011	62%	64%	8%	44% to 72%
2012	63%	64%	9%	39% to 74%
2013	62%	64%	7%	49% to 74%
2015	64%	64%	6%	55% to 82%
2016	63%	65%	6%	51% to 73%
2017	63%	62%	7%	51% to 77%

Note: Estimates represent the percentage of guidelines that were reported across all five journals.

Table 4

Frequency Table of Statistical Models

Factors and Levels in Analyses of Variance (ANOVAs)	Frequency
2-Within	15
2-Within x 2-Within	11
2-Within x 2-Between	7
2-Within x 2-Within x 2-Between	6
3-Within	6
2-Within x 2-Within x 3-Within	4
2-Within x 4-Within	4
3-Within x 2-Within	4
2-Within x 2-Within x 2-Within	3
2-Within x 3-Within	3
2-Within x 3-Within x 2-Within	3
3-Within x 2-Within x 2-Within x 2-Between	3
4-Within	3

Number of Factors in ANOVAs	Frequency
1	29
2	41
3	43
4	26
5	8

Note: #-Within indicates the number (#) of within-subject levels. #-Between indicates the number (#) of between-subject levels (i.e., groups). For the sake of brevity, this table shows the number of factors and levels for those statistical models that were used at least three times in the coded articles, which represents only 72 of the 147 coded articles (49%). Three articles were not included, because they used approaches such as multilevel modelling. All models are shown in the supplementary material on OSF.

Table 5

Numerical Summary of the Relationship between Sample Size, Statistical Power, Large Effect Sizes, and Statistical Analysis

Analysis	Statistical Power	Required Participants	Detectable Effect for Typical *n*
Independent Samples *t* Test	.80	52	.89
	.85	60	.95
	.90	68	1.03
	.95	84	1.14
Paired Samples *t* Test	.80	16	.64
	.85	18	.69
	.90	20	.74
	.95	24	.83
2-Between Group x 2-Within Group Interaction	.80	52	.44
	.85	60	.47
	.90	68	.51
	.95	84	.57

Note: The 'Required Participants' column indicates the number of participants needed to obtain a given level of statistical power to detect a large effect size. A large effect size was considered a Cohen's *d* of .80 for independent samples *t* tests and paired samples *t* test. A large effect size was considered a Cohen's *f* of .40 for the 2-Between Group x 2-Within Group interaction. The 'Detectable Effect for Typical *n*' column indicates the effect size that a study with 21 participants per group and a given level of statistical power would be able to detect. Alpha level was set to .05 for all analyses.

Table 6

Numerical Summary of the Number of Participants Needed to Achieve Statistical Power of .80 for Each Effect Size and Statistical Design

Analysis	Effect Size	Number of Participants
Independent Samples *t* Test	.20	788
	.50	128
	.80	52
Paired Samples *t* Test	.20	199
	.50	34
	.80	15
2-Between Group x 2-Within Group Interaction	.10	298
	.25	50
	.40	22

Note: Number of participants for the independent samples *t* test and interaction effect reflect total number of participants (not number of participants per group).

Table 7

Numerical Summary of the Achieved Power for each Statistical Design, Effect Size, and a Group Sample Size of 21

Analysis	Effect Size	Achieved Power
Independent Samples *t* Test	.20	.10
	.50	.35
	.80	.72
Paired Samples *t* Test	.20	.14
	.50	.59
	.80	.94
2-Between Group x 2-Within Group Interaction	.10	.18
	.25	.73
	.40	.98

Figure Captions

Figure 1. The proportion of guidelines reported across the eight categories of interest.

Figure 2. The proportion of guidelines reported within the following four categories: Participant Information, EEG Recording, Stimulus and Timing, and Preprocessing.

Figure 3. The proportion of guidelines reported within the following four categories: ERP Measurement, Statistics, PCA, and ICA.

Figure 4. These plots show the relationship between statistical power, effect sizes, and total sample sizes for an independent samples *t* test, a paired samples *t* tests, and a test of the 2-Between Group x 2-Within Group interaction in an analysis of variance (ANOVA). Dotted lines represent small (Cohen's $d = .2$; Cohen's $f = .1$), medium (Cohen's $d = .5$; Cohen's $f = .25$), and large effect sizes (Cohen's $d = .8$; Cohen's $f = .4$).

Figure 1

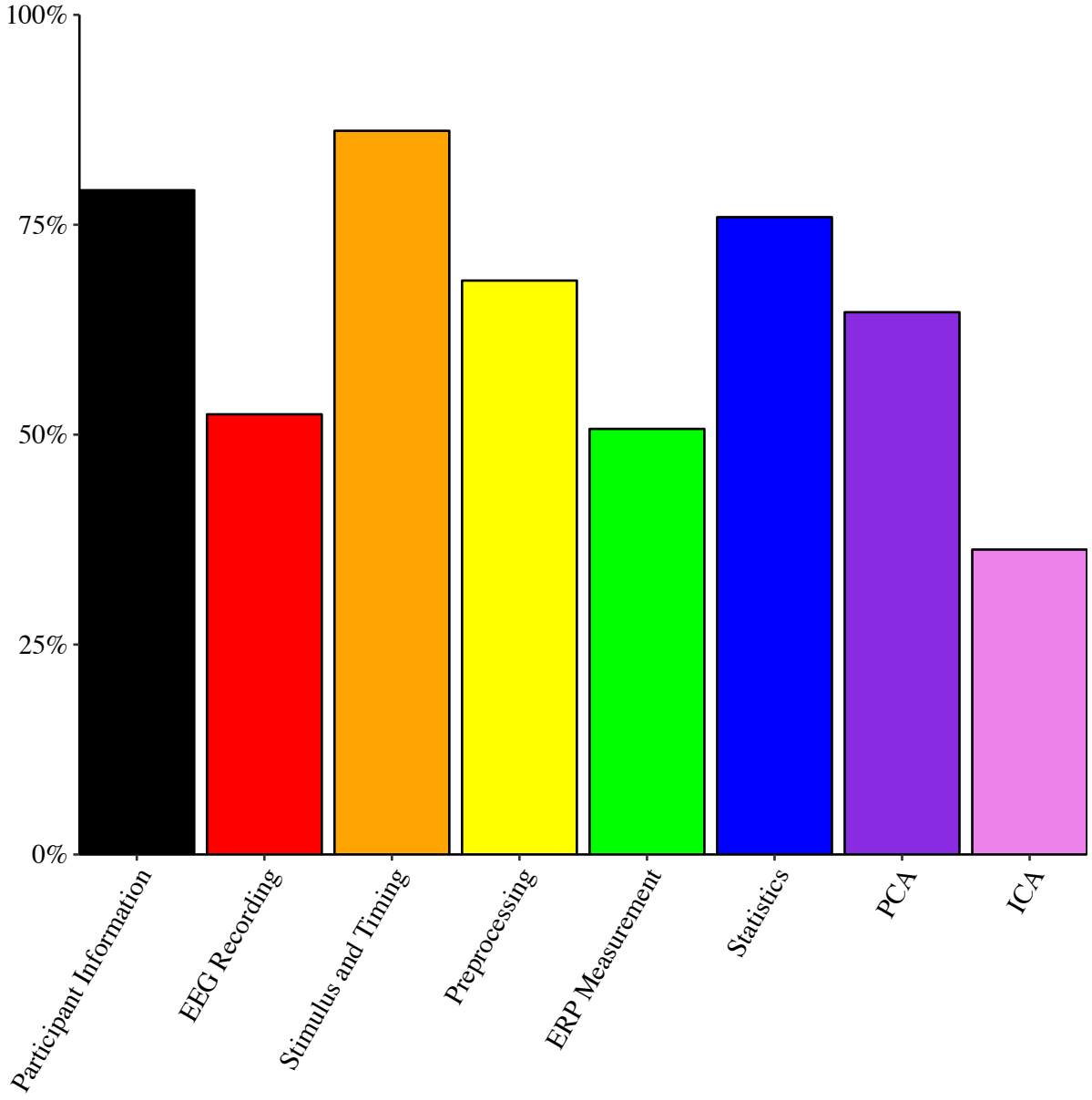

100%
75%
50%
25%
0%
Participant Information
EEG Recording
Stimulus and Timing
Preprocessing
ERP Measurement
Statistics
PCA
ICA

Figure 2

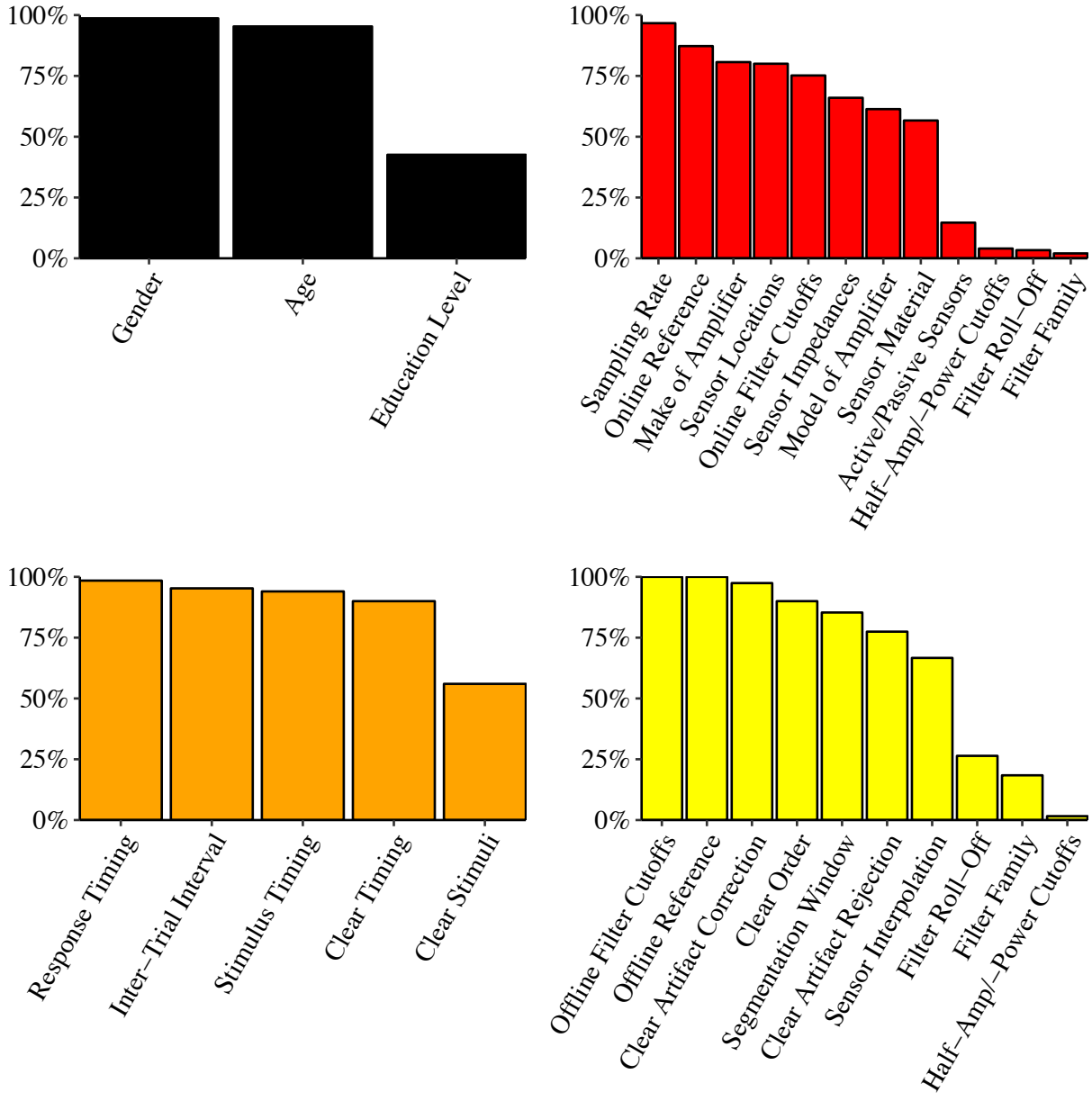

100%
75%
50%
25%
0%
Gender
Age
Education Level
Sampling Rate
Online Reference
Make of Amplifier
Sensor Locations
Online Filter Cutoffs
Sensor Impedances
Model of Amplifier
Sensor Material
Active/Passive Sensors
Half–Amp/–Power Cutoffs
Filter Roll–Off
Filter Family
Response Timing
Inter–Trial Interval
Stimulus Timing
Clear Timing
Clear Stimuli
Offline Filter Cutoffs
Offline Reference
Clear Artifact Correction
Clear Order
Segmentation Window
Clear Artifact Rejection
Sensor Interpolation
Filter Roll–Off
Filter Family
Half–Amp/–Power Cutoffs

Figure 3

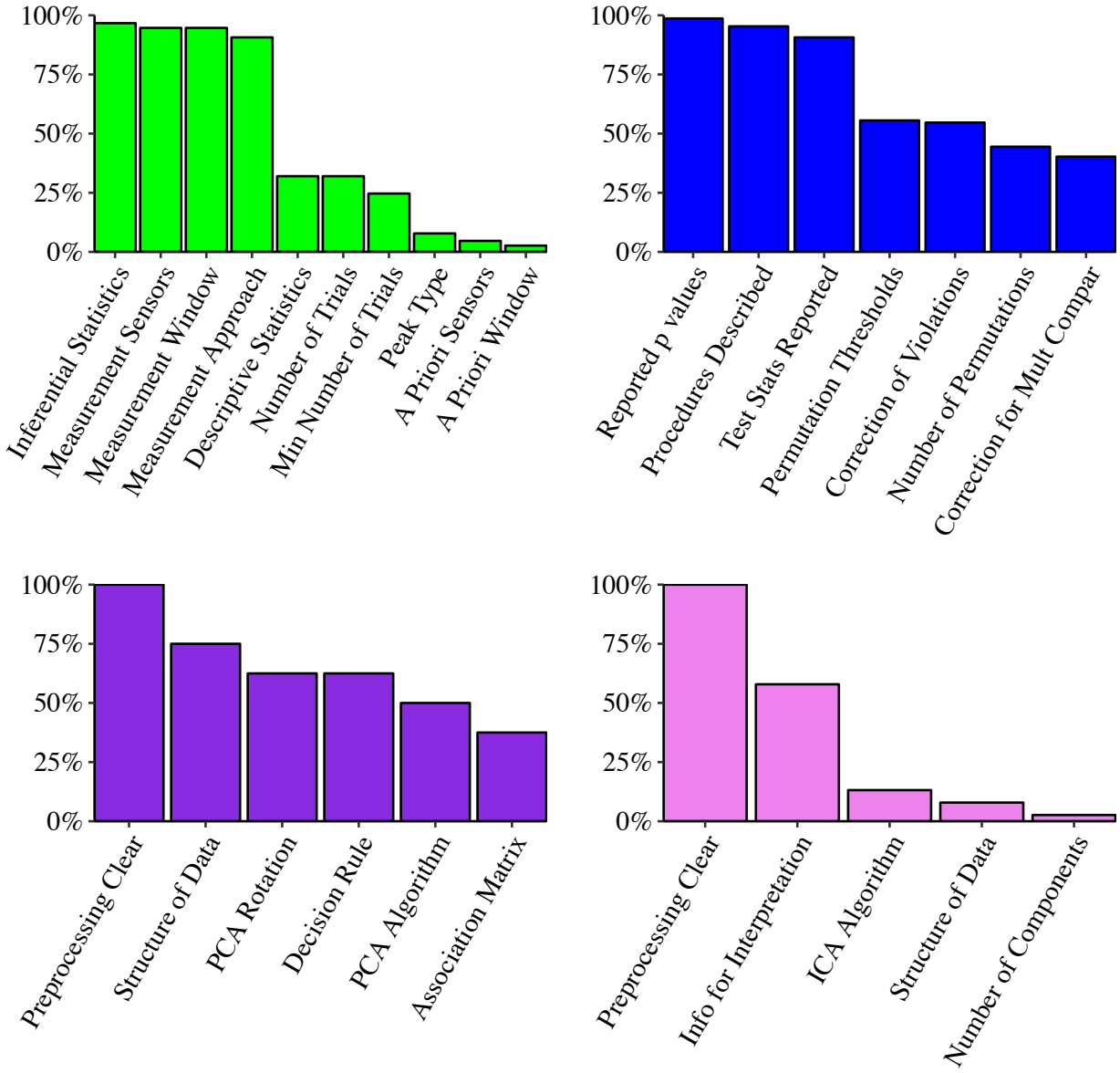
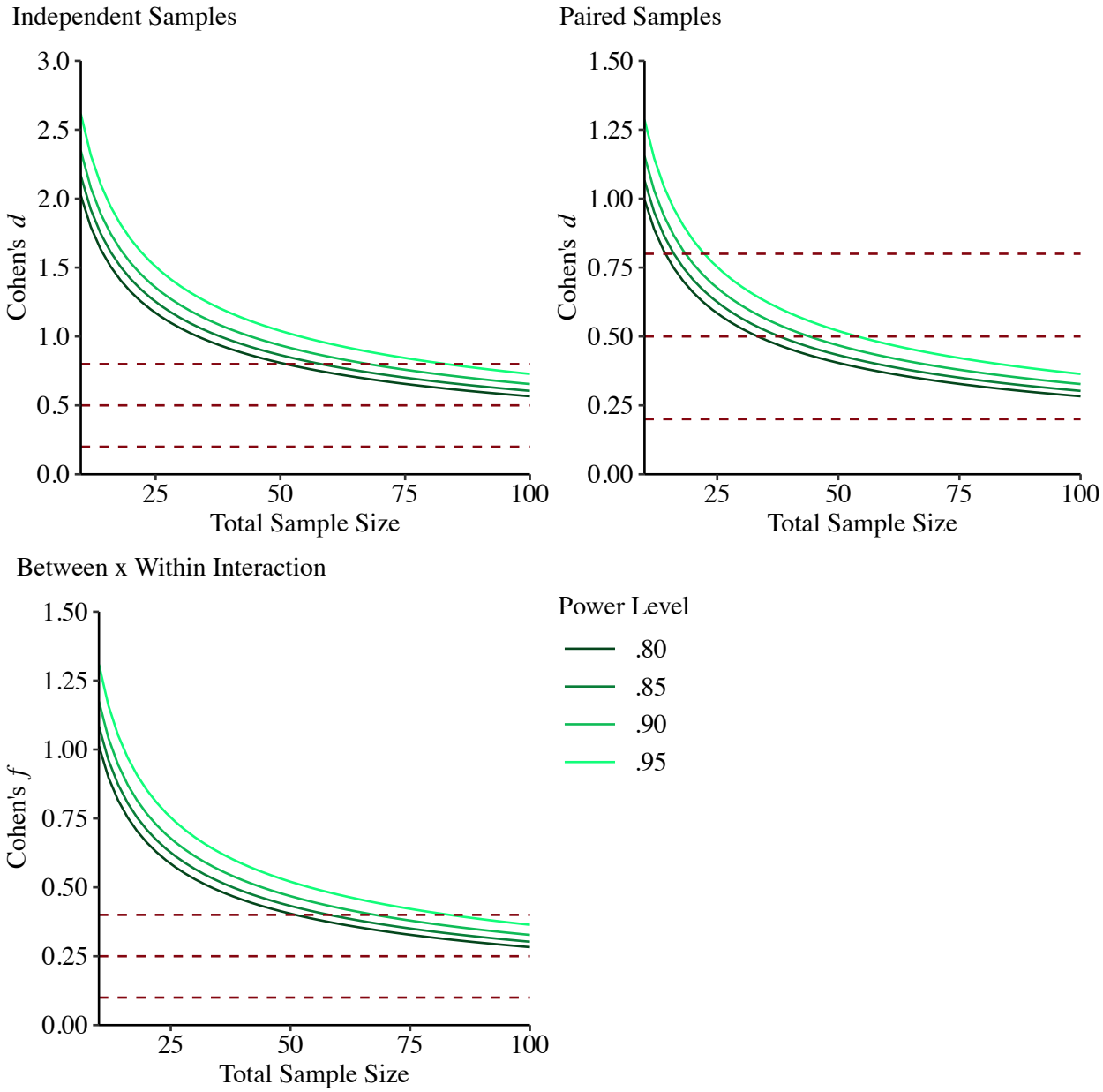

100%
75%
50%
25%
0%
Inferential Statistics
Measurement Sensors
Measurement Window
Measurement Approach
Descriptive Statistics
Number of Trials
Min Number of Trials
Peak Type
A Priori Sensors
A Priori Window
Reported p values
Procedures Described
Test Stats Reported
Permutation Thresholds
Correction of Violations
Number of Permutations
Correction for Mult Compar
Preprocessing Clear
Structure of Data
PCA Rotation
Decision Rule
PCA Algorithm
Association Matrix
Info for Interpretation
ICA Algorithm
Number of Components

Figure 4


Independent Samples
Cohen's d
3.0
2.5
2.0
1.5
1.0
0.5
0.0
25
50
75
100
Total Sample Size
Paired Samples
Cohen's d
1.50
1.25
1.00
0.75
0.50
0.25
0.00
25
50
75
100
Total Sample Size
Between x Within Interaction
Cohen's f
1.50
1.25
1.00
0.75
0.50
0.25
0.00
25
50
75
100
Total Sample Size
Power Level
.80
.85
.90
.95