Abstract

The foundation of scientific progress rests on replicable experiments. Recent investigations have pointed to many issues in the current scientific community: among which we distinguish low powered samples, poor methodological reporting, and high flexibility in research methods. These flaws are despite many efforts including the OSF, the COBIDAS reports, and research journals attempting to encourage better practices. The current project aimed to assess the extent of these issues in EEG research. Randomly sampling 50 research articles published within the last 3 years, we observe that there are many flexible pipelines chosen to analyze data, sample sizes limit power to capturing only large effects, and overall poor methodological reporting limits reproducibility. Overall, these results paint a dim picture: little has progressed in over 5 years ago. We propose solutions to bring forth better practices and hope these findings shed light on the desperate need to enforce standards set up by the scientific community.

Introduction

The scientific method is fruitful and the best tool we have to answer questions empirically. Yet, the process is slow and bound to make many errors along the way. The current prevailing method of statistical inference has attempted to cap the rate of false discoveries at 5%, the infamous $p< 0.05$. This, however, has proven a failure; the rate of false positive is much higher than anticipated, even approximating 40% for the field of neuroimaging (Wager, et al., 2009). Several high-profile attempts of replicating canonical findings in the field of psychology failed (Aarts et al., 2015). This is a cause for major concern, which has loomed over the community for what is now a decade. A growing sentiment of distrust has grown out of this media frenzy, and some call to redefine what we consider significant (Benjamin et al., 2018).

At this crossroads, the scientific community must ask itself what is responsible for this so-called "replication crisis". One culprit is the high degrees of freedom introduced by the researchers themselves—*researcher degrees of freedom* (Simmons et al., 2011). Unintentional p-hacking is commonplace, whereby undisclosed choices researchers make along the scientific process inflate the rate of false discoveries, more than previously anticipated (Simmons et al., 2011; Carp, 2012 a). In a similar vein is method flexibility: fMRI research for example has over 7,000 options to process a single study (Carp 2012). Mathematical simulations and data analysis show that growing flexibility in acceptable methods greatly increases the rate of false positives (Ioannidis, 2005b; Carp 2012a). The number of options will only grow for researchers in the future as novel groundbreaking methods are developed; this concern has no end in sight.

Replication is at the heart of fruitful scientific investigation: it is the engine that drives forward the boundary of knowledge (Moonesinghe et al., 2007). But to replicate findings accurate reports with sufficient details are needed: a foundational principle of scientific communication (Carp, 2012b). Comprehensive methods sections allow us to verify the quality and flexibility of methodologies in a given study and across an entire field. Previous work emphasizes, on the contrary, that methodologies lack important details for replication (Carp, 2012b). This finding is upsetting and greatly impacts how we should think about scientific progress when such a basic tenant is broken.

Taken together, it is important to assess the state of the field in terms of reporting practices to ensure that we can enforce good standards for reproducible science. Given the importance of choices and flexibility in research, the current project aims to describe the extent of preprocessing pipeline flexibility within EEG research under the novel COBIDAS framework developed for MEEG research (Pernet et al., 2018). We hope to answer these burning questions:

1) What is the current state of the field-–how are EEG studies being reported?
2) What areas of scientific reporting could be improved to help reproducibility?
3) How well powered are these studies?
4) How flexible are electrophysiology pipelines?
5) How do these choices influence results and conclusion?

Methods

Article selection

Articles were identified using the PubMed database and selected on the basis of: having the terms "EGG" and "ERP" in their title or abstract; published in English and from 2015 onwards. From the 1075 results, 50 were randomly selected (using a random number generator in R) for further inspection. Of the 50 inspected, any non-experimental research was removed (i.e. meta-analysis, reviews, and all research developing methods, analysis algorithms, new BCI or EEG systems). Thirty-eight articles remained for the analysis. These remaining papers came from an array of 27 different journals, across various impact factors, including: PNAS, PLoS one, Journal of Neuroscience, Cognition, Frontiers, and Psychophysiology.

Article Screening and Criteria

COBIDAS recently released a report on best practices for reporting methods in MEEG (Pernet et al., 2018). To evaluate the quality of the methods reports, a modified version of the COBIDAS criteria was used; we added participant rejection as additional criteria. Due to limited time, only the first three sections of the COBIDAS guidelines were evaluated: *Design and participants*, *Data acquisition*, and *Data preprocessing*. When applicable, supplementary reports were included in the study. When a given article had multiple experiments, the methods of the EEG experiments were evaluated. Whenever parameters could be inferred but were not explicitly stated in an article (e.g. participants were seated and thus their head position can be inferred), information was noted as inferred. Missing information is marked as "NO". In the case that information was present but unclear, it is marked as present but unclear.

Power analysis

To determine the average power of the selected studies from the field, a power analysis was conducted using the "pwr" package in R. Sample sizes, extracted from the above analysis, were used to determine power for varying effect sizes across both t-test and one-way balanced ANOVAS. These tests were chosen as they are the two most common choices of statistical methods in the field. Our power analysis likely reflects a very conservative estimate of power given that many articles had unbalanced two-way and three-way ANOVAs, testing multiple interactions. Notwithstanding, this analysis serves as a benchmark for the field.

Impact of baseline correction

To determine the extent choices affect the conclusions drawn from studies, a small analysis pipeline was setup. Two studies were used to determine the effects of baseline correction on results. Preprocessed ERPs data was used at varying baseline corrections (100ms, 200ms, 250ms, 300ms, 400ms, 500ms) across a sliding window in the peristimulus period (i.e. anywhere from -500 to 0ms before target onset). This allowed us to get a range of p-values and thus assess how a simple choice may affect the conclusions one draws from a given study (see supplementary for detailed methods and code which can be found on github).

Results

Design and Participants

Most papers reported the study population in sufficient detail for replication (see provided excel file) except for the rejection criteria. This could be an error of omission: since they did not remove anyone from the analysis, they did not report it. Yet, omissions give way to assumptions and science is not conducted on assumptions. Moreover, none of the papers described the stimulus presentation computer, nor did 17 papers specify the presentation software, and 24 papers did not specify presentation monitors. None of the studies mentioned publicly available code used for stimuli presentation. In ERP tasks stimulus presentation is fundamental to the design. In addition, 24 studies failed to report the stimuli in detail, sufficient to reproduce the study independently. Taken together, these studies need to improve in two ways: rejection criteria and the behavioral tasks/ stimulus presentation.

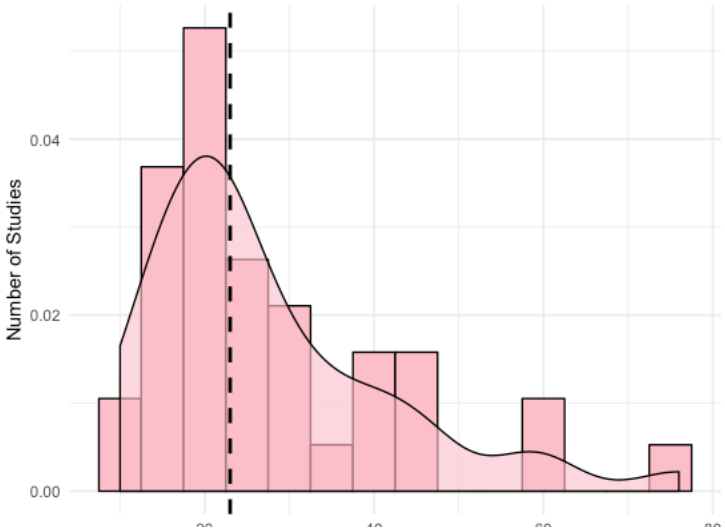

Figure 1. Distribution of sample sizes across all inspected studies.

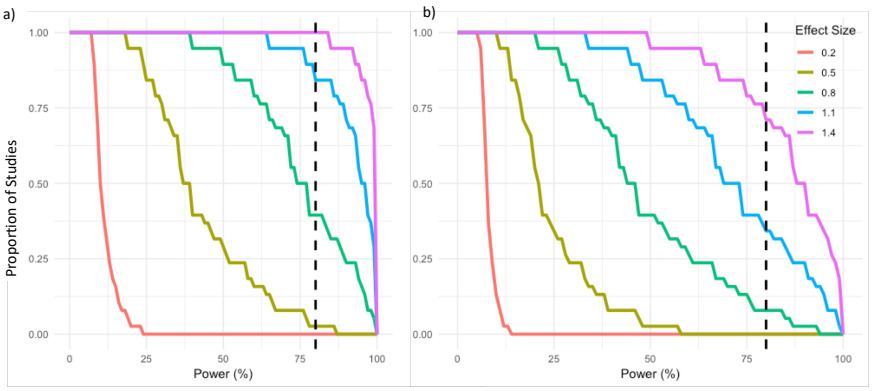

Figure 2. Proportion of studies with enough power at various effect sizes (Cohen's D) for different tests a) Paired t-test b) two sample t-test. For the two sample t-test, sample sizes were divided by two to reflect the power of the study if the study were to be a between groups design, whereas the paired t-test reflects power for within-groups designs.

Power and sample sizes

Neuroimaging studies, specifically fMRI, are significantly underpowered and may only have enough power to determine large effects (Button et al., 2013; Carp 2012b; Poldrack et la., 2017). We wished to characterize the distribution of sample sizes typical of EEG research to then estimate power given common statistical tests. To do so, first sample sizes after exclusions for each study were collected and then entered into RStudio for analysis. Power was estimated using varying range of effect sizes across four types of test: t-test paired, two sample t-test assuming each group would be half of the original sample, a balanced one-way ANOVA with two factors assuming each group was equal to the original sample (essentially doubling the sample size), and finally a one-way ANOVA with two factors assuming each group would be half the size of the total sample. These tests were chosen to explore power in the field as the most commonly chosen

statistic was an ANOVA or a t-test. As seen in Figures 2 and 3, studies are greatly underpowered. This is alarming as power relates to reproducibility of findings (Button et al., 2013). Nonetheless, most studies reported using ANOVAs which as seen in figure 5 suffer even more from low power. These estimates of power are conservative as many reported two-way and three-way ANOVAs tracking various interaction terms. Thus, the field likely can only detect extremely large effects with sufficient power.

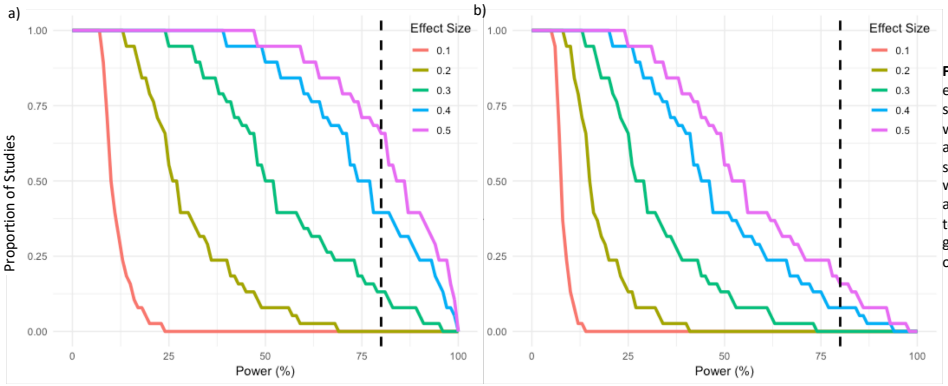

Figure 3. Proportion of studies with enough power at various effect sizes (F) for different tests a) one-way ANOVA with two factors assuming each group had the same size of each study's sample b) one-way ANOVA with two factors assuming each group had half the total number of participants of a given study (sample size / number of factors).

Data acquisition

Several studies failed to report details of the EEG acquisition: electrode material (24 studies), ground (25 studies), electrode and skin preparation (35 studies), recording software (31 studies; none specified the recording computer), 17 did not report impedances (only 5 studies specified when the impendence were checked), and only 5 mentioned stimulus event markers. This is in addition to 24 studies not specifying the recording environment. This is concerning for researchers who wish to reuse this data and those aiming to replicate or reproduce current studies in the field. These key details should be available for future reuse of the data, not to mention these choices likely affect the processing pipelines, data analysis, and conclusions.

Data preprocessing

Lastly, preprocessing methods were examined. Few studies explicitly listed the ordered preprocessing steps were done and assumed one could infer it from how they were presented in the methods. Unspecified details of preprocessing steps include: baseline correction (10 studies), processing software (9 studies), filter types and parameters (21 studies), if they rejected bad channels (29 studies), and how many raters reviewed the raw EEG and segmented trials (38 studies). This uncertainty in methods reporting is accompanied by a great flexibility in processing pipelines including the order in which steps are completed. In short, this reveals the general lack of reproducibility of EEG studies for preprocessing pipelines.

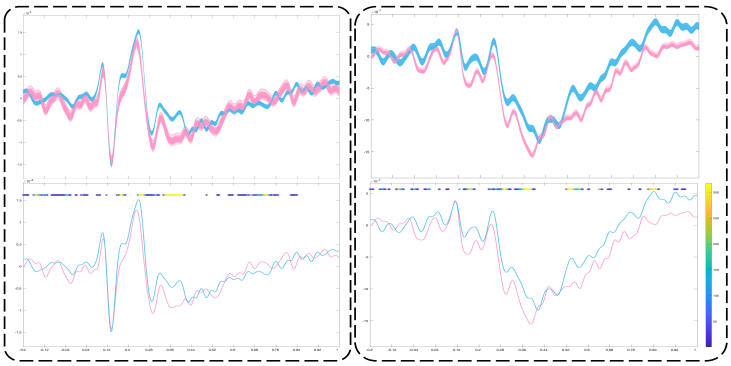

Figure 4. Effects of baseline correction choice on ERP effects (VAN and P300). Top panels depict variation in ERP waveforms given different baseline corrections. Each line depicts a single ERP with a different baseline correction. Bottom panel depicts the significant differences between two conditions (cued and not cued) for all possible baseline corrections as shown above. Stars indicate significance and their colour indicates how frequent this timepoint was significant across all baselines.

Baseline correction

To determine the extent research decisions impact results, we implemented a permutation t-statistic across various choices of baseline correction. As seen in Figure 4 and 5, these choices impact the number of significant results obtained. Fortunately, there are several effects over the neural correlates of consciousness (Figure 4) that surpass statistical significance on every iteration of baseline correction. Yet, there are many smaller effects that are clearly dependent on the choice of baseline correction. This said, not all the choices of baseline correction shown here would necessarily be acceptable: few researchers would choose a baseline from -117 to -17 prior to stimulus onset. On the other hand, Figure 5 stresses how the effects of choice can be drastic, principally in the parietal electrode bundle. If such a simple choice in data analysis could reveal such variability in results, one can only imagine the effects of critical choices concerning preprocessing (e.g. filters and artifact rejection) to even data acquisition parameters (e.g. online filters, reference and even impedances)

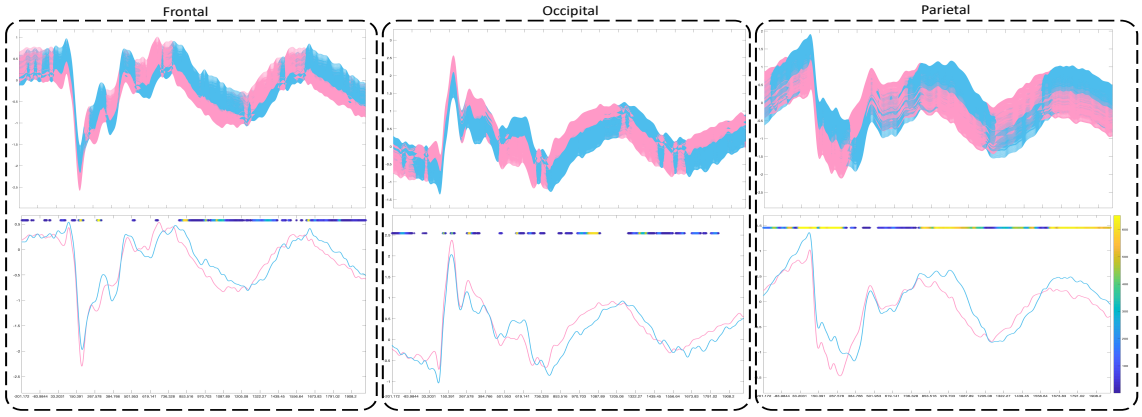

Figure 5. Effects of baseline correction choice on ERPs Top panels depict variation in ERP waveforms given different baseline corrections. Each line depicts a single ERP with a different baseline correction. Bottom panel depicts the significant differences between two conditions (easy--blue and hard--pink) for all possible baseline corrections as shown above. Stars indicate significance and their colour indicates how frequent this timepoint was significant across all baselines.

Discussion

Our present work demonstrates several pitfalls in the field of (but not unique to) EGG research. We firstly describe the extent of research flexibility in pipelines for analyzing ERP paradigms. This works only scratches the surface as oscillatory dynamics are a huge area within electrophysiology. Moreover, we describe the poor, often abysmal, reporting practices across these methodologies: many key details in analyzing data are not mentioned altogether. In addition, the field suffers from a large power-outage: often only having enough power to detect the largest of effects. Lastly, we show how simple methodological choices alter the ultimate results and interpretation. Taken together, these results underline the dire need for change and change is on the horizon. COBIDAS offers researchers a thorough guideline to accurately report their methodologies in great detail for replication. Despite their greatest efforts, it is clear that methodological reporting in EEG has not improved relative to previous reports in similar fields like fMRI (Carp, 2012).

Although these guidelines attempt to simplify the reporting process and balance generalizability across many study designs, I found myself wanting to create a physical checklist for reporting details. This could greatly improve the quality of the research reported and simplify the daunting process of comb through a dense document. The many points of COBIDAS could

be abridged to underline only key details the entire field agrees on. COBIDAS lists compensation, head position, and education as points to include in a study. Arguably there are other key aspects of the study which researchers are failing to report altogether which are more important (event triggers and corrections for timing delays etc.). On the other hand, COBIDAS has omitted rejection of participants altogether. It is common place in EEG to reject participants due to poor signal quality, many artifacts, or poor performance on the behavioral task. This goes beyond the inclusion / exclusion criteria of a study and are important factors to consider. Science relies on random selection and assignment; removing participants from groups is non-random.

The methods reported in these papers reflect a diverse set of subfields in neuroscience, all of which have their own unique methods. Yet, the flexibility in processing pipelines alone is astonishing: almost every paper used a unique pipeline (different parameters, different steps, or a different order). A single cookie cutter pipeline would simply not work; standardization for the entire field is not feasible given the scope of research. By the same token, we need to fully comprehend the extent of the flexibility in the field and study the consequences this has on research outcomes. It is likely that preprocessing pipelines are inherited from research group to trainee and thus the field likely reflects a complex structure of pipelines-inheritance in a pseudo-evolutionary fashion. I would hypothesize that research pipelines have evolved out of trial and error across various groups and cluster within subfield, leading to specific biases.

Proposed Solutions:

1) Create a simple to read and follow **checklist** for researchers to utilize before submitting articles. This may be integrated into the submission process so that standards for methodology reporting are set. Moreover, as it becomes more widely used, it could even be used to inform study design and be integrated early into the scientific process.
2) Intergrade the details of EEG acquisition per COBIDAS into a **standard data format** like BIDS. Recently a BIDS format has been proposed for EEG. An automatic conversion to BIDS format with readable text following the COBIDAS standards could prove indispensable to the filed bettering reporting standards and data sharing.
3) **Fix COBIDAS**. As mentioned, participant rejection is an area we feel COBIDAS could improve upon. A ranking of their criteria could reflect the most vital (non-negotiable) aspects needed to be reported thus limiting the length of the list and addressing the most pressing issues for all researchers (regardless of research topic).
4) Use COBIDAS as a **rating system** to rate papers in terms of their methodological reporting and a theoretical reproducibility scroe. This could then be used by others as targets for reproducing findings, and metrics for journals to achieve (i.e. our papers in our journal have the highest RePRO factor)
5) **Incentivize researchers** to report methodologies in detail. Completed checklists could be implemented in a machine-readable way such that this would generate an automatic methods section saving researchers time. Moreover, pre-registered reports could be used to fill methodological checklists and documentation for open-data /code such that researchers are incentivized to pre-register, complete methods checklists, and share their data/code in a one step-process.
6) Make **methods of papers machine readable** across journals. This could aid in the meta-study of research as well as serve as a self-check system whereby technologies could then estimate the flexibility in research methods, the typical pipelines used, and consensus on analysis in the field. Machine-readable methods for studies could be enforced by journals

such that their infrastructure would offer researchers tools to not only better report their methods, but also tools to study the direction in which fields are headed. This, for example, could lead to large scale estimates of power.

Conclusion

Taken together, the results from this study demonstrate three key factors about EEG research: 1) there is a great diversity and flexibility in processing pipelines; 2) Many studies fail to report key methodological details for task design, EEG acquisition, as well as preprocessing; 3) the sample sizes of these studies are small—having power to detect only large effects. Our findings paint a grim picture of the current status of research. Notwithstanding, there have been many great efforts to improve the current state of the field (e.g. COBIDAS, OSF, and open data initiatives) which have yet to take off. A widespread mindset shift is needed in science. To benefit from scientific progress, we must be willing to adapt our current patterns to steer the community towards better ones—research practices that we know will produce better *reproducible* findings.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., ... Zuni, K. (January 01, 2015). Estimating the reproducibility of psychological science. *Science, 349*.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

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*, 365.

Carp, J. (January 01, 2012 a). On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in Neuroscience, 6*.

Carp, J. (October 15, 2012 b). The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage, 63,* 1, 289-300.

Ioannidis, J. P. A. (January 01, 2005). Why Most Published Research Findings Are False. *Chance, 18,* 4, 40.

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (February 01, 2007). Most Published Research Findings Are False—But a Little Replication Goes a Long Way. *Plos Medicine, 4,* 2.)

Pernet, C. R., Garrido, M., Gramfort, A., Maurits, N., Michel, C., Pang, E., … Puce, A. (2018, August 9). Best Practices in Data Analysis and Sharing in Neuroimaging using MEEG. https://doi.org/10.31219/osf.io/a8dhx

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (November 01, 2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22,* 11, 1359-1366.

Van Noordt, S. J. R., Desjardins, J. A., & Segalowitz, S. J. (January 01, 2015). Watch out! Medial frontal cortex is activated by cues signaling potential changes in response demands. *Neuroimage, 114,* 356-370.

Wager, T., Lindquist, M., Nichols, T., Kober, H., Van Snellenberg, J., 2009. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. NeuroImage 45, S210–S221.