International Journal of Psychophysiology 111 (2017) 1–4

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Rigor and replication: Toward improved best practices in human electrophysiology research

There is an increasing focus on transparency, rigor, and replication in psychological science and science in general (e.g., Ioannidis, 2014; Larson, 2016; Open Science Collaboration, 2015). The emphasis on clear and replicable science comes on the heels of failed replications of studies that have served as the foundation of multiple theories, challenging what is known about many long-standing psychological phenomena (Hagger and Chatzisarantis, 2016; Open Science Collaboration, 2015). Across fields of science, studies adequately powered to detect the effects of interest remain elusive–thus, many published effects likely represent inflated estimates (Bezeau and Graves, 2001; Cashen and Geiger, 2004; Chan and Altman, 2005; Maggard et al., 2003). Indeed, low statistical power may undermine conclusions in much of the extant psychological, neuroscience, and medical literatures (Button et al., 2013; Ioannidis, 2005, 2008). Poor replication and less-than-reliable scientific practices have the potential to erode the public trust and perception of research findings, increasing the already difficult task of dissemination and implementation of research findings. Questionable scientific practices and difficulties with replication are also associated with an increasing concern that false findings may represent the majority of published research (Ioannidis, 2005).

Inability to replicate prominent studies along with a spotlight on statistical insufficiencies, methodological problems, and even fraudulent scientific reports (e.g., Broockman et al., 2015) have led to what is being termed as a "replication crisis" in both the field of psychology and science more broadly. Psychophysiological research, and human electrophysiology more specifically, has not yet been specifically targeted as an area with poor replication; however, there is no reason to believe the field of psychophysiology is an exception to the difficulties proliferating scientific research (Baldwin, 2017–in this issue). Indeed, the incentive structure in academia that prizes frequent publication over rigor (Baldwin, 2017–in this issue; Cohen, 2017–in this issue; Nosek et al., 2012), the presence of a high level of researcher flexibility that may increase Type I error (Simmons et al., 2011), the relatively understudied psychometrics of psychophysiological measurements (Baldwin et al., 2015; Clayson and Miller, 2017–in this issue-a), and the existence of small sample sizes without *a priori* sample size calculations (Guo et al., 2014; Larson and Carbine, 2017–in this issue) among other difficulties, suggest striking similarity between problems throughout science and psychophysiological research.

One way to improve the rigor and replicability of science generally, and human electrophysiology science specifically, is a renewed focus on best practices and increasing openness and transparency in the research design, data collection, and data analysis pipeline. Thus, this special issue was designed with an eye toward communication and implementation of improved best practices in human electrophysiological research—specifically electroencephalogram (EEG) and event-related potential (ERP) research often submitted to the *International Journal of Psychophysiology* (IJP). Whereas human electrophysiology is the focus of this special issue, many of the recommendations and best practices provided here are relevant across methods and areas of research. We are grateful for contributions from prominent scientists focused on EEG and ERP research who, rather than simply focusing on identifying insufficiencies, provide ways to improve psychophysiological research and put forward best practices and methodological suggestions that will increase transparency, improve reliability, and ultimately improve replicability in our field. The papers in the special issue provide guidelines applicable to both novice and seasoned researchers, examples of replicable research, and (in many cases) the authors have provided checklists, guidelines, code, or toolkits researchers can use to improve their labs going forward.

The special issue begins with a theoretical piece written by a methodologist/quantitative psychologist, Scott Baldwin, who provides an insightful look at human electrophysiological research from a methodologist's standpoint. Specifically, Baldwin (2017–in this issue) identifies roadblocks he has seen to rigorous psychophysiological research, including misplaced incentives for publication above rigor, researcher flexibility, low statistical power for firm conclusions, and poor precision in measurement. He then provides five suggestions for improvement including increasing statistical power through collaboration between labs, improving statistical and methodological training in our graduate programs (as an aside, one surprising finding from Baldwin's analyses is that neuroscience programs have fewer methodology/statistics courses than social psychology programs where much of the replication literature is currently focused), pre-registering studies, improving reporting standards in our research, and shifting incentives in our institutions from publication to high-quality and methodologically-rigorous science. Although many of these suggestions are common across areas of science, Baldwin provides several examples of particular relevance to psychophysiologists.

Moran et al. (2017–in this issue) follow with a tutorial on meta-analysis and a demonstration of its utility by examining relationships between action monitoring ERPs and depression. Meta-analysis represents a useful tool for evaluating the robustness of accumulated effects in an attempt to arrive at a consensus regarding a body of work, which, in the case of human electrophysiology, is often based on primary studies comprised of small samples and varied methodologies. The robustness of the findings can then be used to arbitrate extant theories

http://dx.doi.org/10.1016/j.ijpsycho.2016.12.001

or inform the practical relevance of the research. Moran et al.'s step-by-step tutorial is broad in nature so as to be useful to many scientists, but specifically focuses on the application of meta-analysis to neurophysiological findings. The general 5-step process – formulating the problem, conducting the literature search, coding studies and extract data, synthesizing effect sizes and assessing for heterogeneity, and assessing for threats to validity – provides a comprehensive and straightforward roadmap for researchers to investigate the strength of a research base. Each step is demonstrated using results from the literature on the relationship between action monitoring ERPs – the error-related negativity (ERN) and feedback negativity (FN) – and depression. This paper, therefore, combines guidelines for conducting robust meta-analyses with novel empirical results. The results of the novel meta-analysis demonstrate that publication bias likely contaminates the literature on the relationship between the ERN and depression and that the relationship between the FN and depression is dependent on task demands. Moran et al. conclude with recommendations for neurophysiologists of primary studies to aid in the development of a transparent and informative research base from which robust meta-analyses can draw. Perhaps most importantly, these recommendations call for authors of primary studies to report adequate descriptive and effect size statistics, including the full results from non-significant effects that are often buried in summary statements (e.g., "correlations between ERN and behavioral measures were non-significant, $rs < 0.10$, $ps > 0.20$.").

One frequently noted recommendation to improve research practices is to complete *a priori* sample size calculations to ensure that studies are sufficiently powered and to safeguard against inflated effect sizes (Button et al., 2013; Ioannidis, 2005, 2008; Sawyer and Ball, 1981; Sterne and Smith, 2001). Larson and Carbine (2017–in this issue)randomly selected 100 clinically-relevant EEG/ERP studies to identify the frequency of use of *a priori* sample size calculations and if the necessary information is being presented in EEG/ERP studies to make such calculations. Findings from this study are sobering, as 0 out of the 100 studies provided an *a priori* sample size calculation. It is possible that such calculations were made, but not reported; however, an emphasis on how sample sizes were determined is clearly needed in human electrophysiology research. Furthermore, only 40% of studies reported effect sizes, with 56% reporting mean values and 47% reporting variance values such as standard deviations or standard errors. One clear consequence of underpowered studies is large variability in published research findings, such as the variability that is often seen in the ERP literature (Maxwell, 2004). Following the reporting standards set forward by Keil et al. (2014) that emphasize reporting of sample sizes (or any number of reporting guidelines) along with the information needed to conduct sample size calculations is clearly needed.

In the Larson and Carbine paper, they indicate that 77% of their sample used repeated-measures statistics. Given the prevalence of repeated-measures statistical analyses, Joe Dien (2017–in this issue) provides an important best practices paper for using repeated-measures ANOVAs in ERP research. The paper uses specific analyses of a previously-published dataset to show that a robust implementation of repeated-measures ANOVA (i.e., robust ANOVAs) can protect against spurious findings, even with a slight loss of statistical power. Specific suggestions for optimal settings for robust ANOVAs including the number of bootstrap runs, determining the variability of *p*-values and reducing the alpha threshold, and the degree of outlier trimming are provided. The analyses further indicate that regional channel groupings (e.g., averaging several channels into a single region-of-interest [ROI]) in ANOVA designs improved noise levels, but diluted overall effects. These suggestions should be considered for accurate analysis of repeated-measures designs.

Utilizing appropriate statistical analyses under the correct assumptions and having precise measurements with reported psychometrics are critical for accurate and rigorous research. Clayson and Miller (2017–in this issue) provide a *tour-de-force* review of the importance of both measuring and reporting accurate reliability (or in the case of generalizability theory [g-theory], dependability) information in ERP research. Specifically, they emphasize that the psychometric properties of ERPs are context dependent and, therefore, should be evaluated and reported for each study. They provide guidelines for improving psychometrics in ERP research, including reporting the threshold for acceptable reliability for each study (above 0.70 or 0.80 are recommended), specifying how reliability for the study was calculated, reporting the reliability estimate for the true scores, and justifying how the minimum number of trials (and reporting this number) was determined.

Notably, Clayson and Miller (2017–in this issue-b) did not stop with simply providing recommendations, they provide a second paper wherein they outline an open-source toolbox, called the ERP Reliability Analysis Toolbox, for calculating g-theory-based dependability measures for ERP studies. The toolbox is free, available on-line, and provides a valuable tool in determining measurement reliability that can be easily used and implemented by investigators throughout the field of ERP research. The strengths of the g-theory approach to psychometrics, including not requiring the same number of trials per measurement (i.e., unbalanced designs are acceptable), providing reliability estimates for multiple facets in the same analysis, and not requiring parallel forms are outlined along with directions for use, developmental procedures, and the value of the ERP Reliability Analysis Toolbox. We feel the open-access nature of this toolbox is an excellent contribution to neurophysiological science.

The areas of time-frequency and asymmetry analyses are growing with increased novelty, sophistication of findings, and data analysis approaches. Thus, the next three papers provide best practice guidelines and critical information for the utilization of time-frequency and EEG asymmetry data. First, Mike X. Cohen (2017–in this issue) reviews ways to improve rigor and replication in time-frequency analyses. He starts with an excellent review of why replication is important throughout cognitive electrophysiology research, then provides specific recommendations for improving the probability of replication in time-frequency analysis. Recommendations include utilizing appropriate experimental design, utilizing a sufficient time period for baseline normalization, ensuring sufficient numbers of trials and participants, including data and code in a transparent and open manner, pre-registering study design and analysis plans, and emphasizing the publication of null results (among other excellent recommendations). Cohen's review would be of benefit to both novice and expert researchers conducting time-frequency analyses.

Although time-frequency analyses allow for the estimation of the magnitude of regional brain activity, more recently, time-frequency phase-synchrony (TFPS) measures have been developed to capture functional connectivity between brain regions. This is where Aviyente et al. (2017–in this issue) pick up the discussion by reviewing novel methods they have worked on to quantify functional connectivity from EEG and ERP recordings. They begin by discussing the emergence of EEG-based functional connectivity measures, as investigators examining traditional fMRI functional connectivity became dissatisfied with its temporal resolution to capture dynamic interplay between regions over short timescales. As EEG recording arrays became denser over time, the promise of using EEG and ERP TFPS to index functional connectivity was realized. The heart of their paper includes a critical review of the numerous means by which TFPS can be calculated to index functional connectivity and a demonstration of the superiority of a novel measure they have developed particularly for ERPs. The demonstration includes a new set of analyses showing that their method nicely captures increased connectivity between medial and lateral prefrontal regions during inhibitory and loss feedback processing. In all, their paper highlights the exciting potential for ERP-based functional connectivity measures to reveal rapid dynamical interplay between brain regions.

The Smith et al. (2017–in this issue) paper represents a transition from a focus on cognitive mechanisms to the use of frequency-band