International Journal of Psychophysiology 111 (2017) 5–16

Contents lists available at ScienceDirect

International Journal of Psychophysiology

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

Improving the rigor of psychophysiology research

Scott A. Baldwin *

Department of Psychology, Brigham Young University, United States

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form 11 April 2016
Accepted 20 April 2016
Available online 29 April 2016

Keywords:
Replication
Rigor
Statistical power

ABSTRACT

Psychology as a field is in the midst of what is sometimes called a "crisis" because false findings are prevalent. Although the focus of the methodological and substantive criticisms of psychology has focused on social psychology, psychophysiology research is not without its problems. The author discusses (a) researcher flexibility and its impact on the stability of conclusions and (b) the role power plays in the probability that a finding is true and the precision of estimates. The author uses examples and data from psychophysiological research to illustrate the problems. The author concludes with a discussion of ways to shift the practice of science to improve the reliability of findings. Suggestions for improvement include: increased power through collaboration, improved statistical and methodological training, pre-registration of studies, improved reporting standards, and shifting incentives surrounding hiring and promotion.

> *There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims* (Ioannidis, 2005, p. 0696)

1. Introduction

Rigorous, replicable, thoughtful, transparent, correcting—these are all adjectives used to describe science. Research in psychology, neuroscience, medicine, and other disciplines have been said to be in "crisis" of late (Pashler and Harris, 2012), where the word crisis is used to denote that some of the fundamental assumptions we often (tacitly) make about the scientific literature may not be true. False findings abound (Ioannidis, 2012; Ioannidis and Trikalinos, 2005). More instances of conscious fraud have been exposed (e.g., Bhattacharjee, 2013; Bohannon, 2015); efforts to replicate studies have demonstrated that many findings have not yet been replicated (Open Science Collaboration, 2015); and major journals published studies making improbable claims (Bem, 2011; Wagenmakers et al., 2011). This article aims to briefly review some of the key issues that cast doubt about the conclusions drawn in the psychophysiology literature, with a particular focus on the electrophysiology literature, as well as review potential methods for improving the quality of the literature. My comments are from the prospective of a methodologist who collaborates with researchers in psychophysiology as well as other sub-disciplines in psychology. I also consider how the incentives in academic culture create a context in which change to publishing practices is difficult.

1.1. What counts in academics

Considerations about improving the rigor of research cannot ignore the culture of academia, specifically how that culture shapes the day-to-day activities of scientists. Those activities may seem a bit bizarre to the casual observer, and, if not bizarre, at least unexpected. When I started college my image of a scientist was someone working in a lab, recording data, and running experiment after experiment trying to uncover a truth about the world. I was not aware that scientists often have a team of people that collect and analyze data and that scientists spend much of their time writing papers and grants. Further, I was not aware that one's standing as a scientist largely comes from the number of papers one produces, how much attention those papers engender, and how much money one brings to the university.

In the book *Laboratory Life: The Construction of Scientific Facts*, Latour and Woolgar (1986) document a two-year anthropological study of scientists. Latour spent two years at the Salk Institute in the late 1970s, observing the work of scientists as would an anthropologist observing a previously unknown tribe. Latour strove to take an "outsider's" view of the work, aiming to make sense of varied behaviors, rituals, and norms he observed. After observing the day-to-day work of the scientists, Latour and Woolgar (1986) write:

> The production of papers is acknowledged by participants as the main objective of their activity. The realisation of this objective necessitates a chain of writing operations from a result first scribbled

* Corresponding author at: Brigham Young University, 268 TLRB, Provo, UT 84602, United States.
E-mail address: scott_baldwin@byu.edu.

http://dx.doi.org/10.1016/j.ijpsycho.2016.04.006
0167-8760/

> on a sheet of paper and enthusiastically communicated to colleagues, to the final registering of published literature in the laboratory archives. The many intermediary stages (such as talks with slides, circulation of preprints, and so on) all concern literary production of one kind or another. (p. 71)

After noting the extensive time and resources that are involved in producing these papers, Latour and Woolgar wonder: "…how can a paper be both so expensive to produce and yet so highly valued? What exactly can justify participants' faith in the importance of the papers' contents?" (p. 71–72).

Scientists are judged by their vita and adding publications, especially publications in prestigious journals, is how careers are made. Consider the advice given to new psychology graduate students about the hiring process in *The Compleat Academic*, a popular book aimed at mentoring graduate students and new faculty entering academia:

> The information we need to arrive at a short list of applicants is contained in the letters of recommendation and, primarily, in the academic vita. Wise graduate students, therefore, will start at day one of their first year in a PhD program to develop a strong vita….Alter your perspective so that you derive your professional self-respect entirely from what is on that document. From the start of graduate school on, throughout what we hope will be a long and productive career, you *are* your vita. (Lord, 2004, p. 10, emphasis in original)

I suspect most academics reading this quote would nod in general agreement. To be sure most psychologists want to learn about the world and human behavior. However, the realities of hiring, tenure, grants, and awards, where lines on a vita are paramount to success, are strong and influential. We may not like being reduced to a vita, but, like the scientists observed by Latour and Woolgar, we spend a lot of time producing papers and our careers are judged by those papers.

Publishing papers is not a problem; articles and books are the primary method for scientific communication. Nor is seeking accolades by definition a problem. Science is not a zero-sum game where one is either pursuing knowledge or accolades. However, when one's job is on the line or one is aiming to be the first to publish in a particular area, producing reliable, replicable knowledge may not be as important as producing publishable knowledge. Likewise, when pursuing grant money, questions that are fundable may take priority over questions that are most theoretically relevant.

1.2. Aims

Below I discuss the consequences of researcher flexibility and underpowered studies on the quality of research findings. Although these issues, and the recommendations for how to address them have discussed a lot recently, none is new. We have known about the problems of power, reliance on *p*-values, and excessive researcher flexibility for a long time (e.g., Cohen, 1962, 1994). However, research practices have not changed; we rely on *p*-values as much as ever and we continue to publish underpowered studies. I suspect that some of this "cultural inertia" regarding research practices is associated with how we define success and prestige in academics. Most suggestions for improving the rigor of research lead to fewer publications, more null findings, and more transparency regarding research practices. These reduce the size of our vita, and thus, according to *The Compleat Academic*, our identity as researchers.

The primary aim of this paper is to review some of the latest trends in the methodological literature regarding (a) roadblocks to rigorous research and (b) strategies for improving rigor. Specifically, I discuss (a) researcher flexibility and its impact on the stability of conclusions and (b) the role power plays in the probability that a finding is true and the precision of estimates. I conclude with a discussion of suggestions for addressing these problems. I discuss these recommendations in the context of academic culture because change is not likely to happen without consideration of the day-to-day context in which research occurs.

2. Does psychophysiology need to improve?

I noted previously that the problems regarding replication, rigor, and fraud have been called a "crisis". One might argue that the attention these problems receive is overblown or, at the very least, the problems are limited largely to social psychology and to some extent fMRI. After all, the focus of the Replication Project was social and cognitive psychology (Open Science Collaboration, 2015) and only one study was clearly within the psychophysiology area (Hajcak and Foti, 2008).[1] Psychophysiology may not be similar to these sub-disciplines and may have more robust findings.

This reasoning is problematic because it assumes that psychophysiology is an exception to the problems common to social psychology and other sub-disciplines. Psychophysiology and neuroscience students do not typically receive more methodological training during graduate school than students in other areas. A survey of the top 50 *U.S. News and World Report* psychology programs with a neuroscience degree (36 responded to the survey), showed that neuroscience students are required to take fewer methodology classes than social/personality students (Schwartz et al., 2016). Measurement in psychophysiology is not particularly rigorous. Although psychophysiology measures often are seen as hard evidence because they are measures of physical phenomena, these measures are not particularly strong and sometimes struggle to meet standards of reliability. For example, in fMRI, average reliability is approximately 0.5, averaged across a number of measures of reliability (e.g., test–retest, voxel counts) and tasks (Bennett and Miller, 2010). Two studies have shown that in EEG studies of error-related negativity (ERN), the number of trials needed to obtain reliable estimates exceeds what is often used (Baldwin et al., 2015; Larson et al., 2010). Although there is some disagreement regarding the number of trials needed in ERN research (Foti et al., 2013; Meyer et al., 2013; Olvet and Hajcak, 2009; Pontifex et al., 2010), what is clear is that reliability of these EEG measures is not firmly established.

As I discuss below, statistical power is major issue for the replicability and quality of research findings and low power plagues studies across psychology (Button et al., 2013; Cohen, 1962; Open Science Collaboration, 2015). I am unaware of any evidence that suggests psychophysiology as a whole is well-powered and I present some evidence that suggests just the opposite.

Perhaps psychophysiology research is notably replicable? This is an empirical question and there is no clear evidence to answer one way or another. The replicability project included one psychophysiology study (Hajcak and Foti, 2008). A key finding reported in this paper was that the magnitude of the ERN was negatively correlated with startle responses—making errors is associated with increased startle (Hajcak and Foti, 2008). This correlation was not significant in the Replication Project (https://osf.io/jret9/) nor have the original authors replicated the finding (Riesel et al., 2013). Finally a re-analysis of the data from Hajcak and Foti (2008) indicated that the significant findings disappeared after excluding a single outlier (Moser et al., 2014). This was just one study and one finding. Perhaps this would be unique in psychophysiology research. However, without clear evidence that psychophysiology is particularly rigorous with respect to the design and analysis features that lead to high probability of replication and impact, it is reasonable at this point in time to encourage psychophysiology researchers to increase rigor.

[1] Scholars have debated the merits of the Replication Project (Anderson et al., 2016; Gilbert et al., 2016). Although the Replication Project is not without its problems, it is one part of the broader examination of the problems in psychological research discussed above. Consequently, one does not need to rely on the results of the Replication Project to make the case that much psychological research could be more rigorous.

3. Researcher flexibility

Among the key concepts in any introductory statistics course in psychology and the behavior sciences in the concept of Type I errors (Pagano, 2013). That is, rejecting the null hypothesis when it is in fact true. Type I errors are scientifically problematic because they provide false support for theories. To reduce Type I errors, we set an α-level, which represents the long-run rate of Type I errors we will tolerate and is most often set to 5%.

A number of scholars have demonstrated that researcher flexibility in the design and analysis of studies can inflate the Type I error rate above α, regardless of the *p*-value associated with the statistical test (Gelman and Loken, 2014; Ioannidis, 2005; Simmons et al., 2011). This flexibility has been termed *researcher degrees of freedom* (Simmons et al., 2011), *The Garden of Forking Paths* (Gelman and Loken, 2014), or just plain *bias* (Ioannidis, 2005). Regardless of what flexibility is called, it refers to design and analysis choices that increase the probability that a research finding will be judged to be statistically significant and thus increases the probability that a finding will be published (Greenwald, 1975).

Simmons et al. (2011) describe some of the decisions that researchers make: "In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?" (p. 1359). Other decisions include: How should missing data be handled? Should the analysis method be what reviewers and editors expect (rather than methods that best accommodate the design and data types used)? Which interactions should be tested? Should analyses be added at the suggestion of a reviewer, even if the analyses are exploratory? Should null-hypothesis tests be one-tailed or two-tailed?

Most of these choices are present in any study. If the choices were all made and committed to prior to the collection of data, then α-levels would not be affected. The trouble is that researchers often make these choices after having seen the data, making *p*-values contingent upon the data—contingent upon this specific dataset in this particular study. However, *p*-values are long-run probabilities that involve many possible datasets. Gelman and Loken (2013) note: "Once we recognize that analysis is contingent on the data, the *p*-value argument disappears—one can no longer argue that, if nothing were going on, that something as extreme as what was observed would occur less than 5% of the time" (p. 12).

Given the many, often dataset specific, decisions are made in any study, this process has been termed: "The Garden of Forking Paths" (Gelman and Loken, 2014). Fig. 1 illustrates the forking paths metaphor. The top of Fig. 1 begins with a study with multiple conditions as well as possible outcomes of a manipulation. Fig. 1 follows the forking paths through some of the possibilities in which researchers decide whether to throw out conditions where the manipulation failed, whether to use one- or two-tailed tests, whether to report all outcomes, whether to drop outliers, and whether to explore potential interactions (Simmons et al., 2011). As researchers make their way through these decisions, the resulting *p*-values are dependent on the unique aspects of the data in their particular study, making the interpretation and utility of the *p*-value unclear (Gelman and Loken, 2014).

I suspect most researchers do not simply go on a fishing expedition with their data (Gelman and Loken, 2013). However, when one combines the incentives to publish (see above) with the fact that significant results are easier to publish than non-significant results, it is not surprising that researchers make choices that maximize the chances of obtaining $p < 0.05$. Consider the advice from Bem (2004) about writing a journal article:

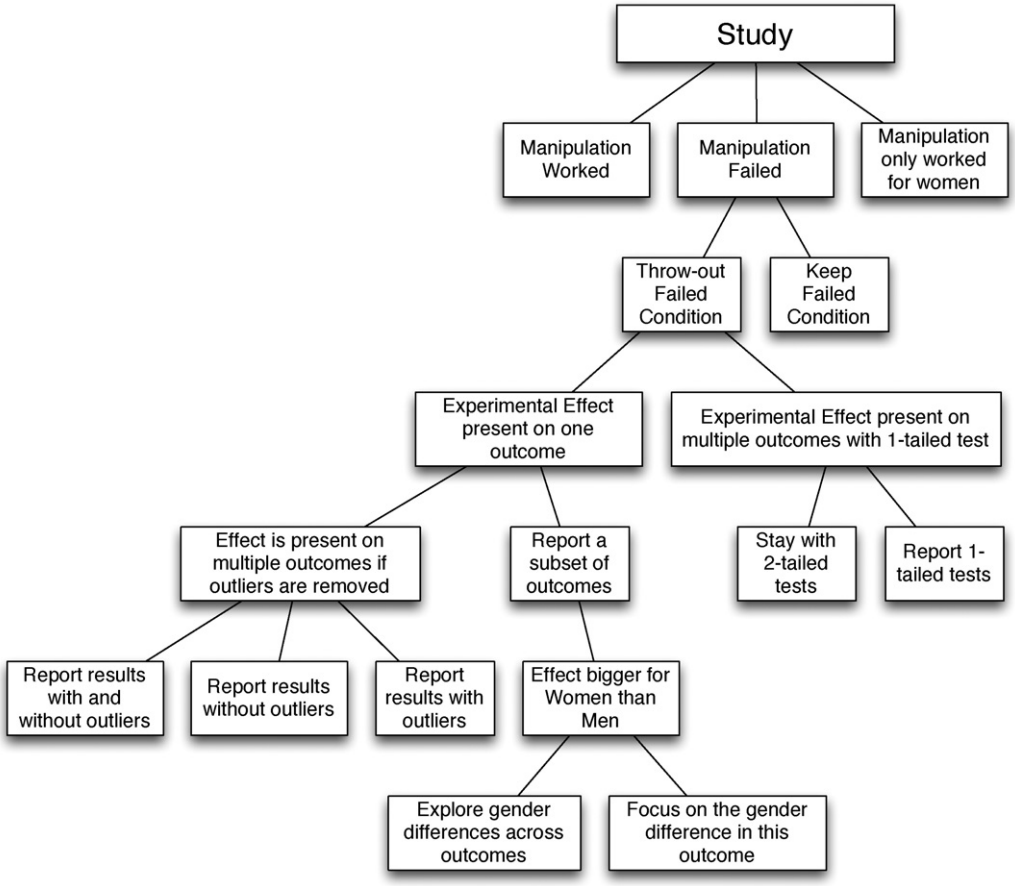

Fig. 1. Illustration of design and analysis decisions in The Garden of Forking Paths.

> To compensate for this remoteness from our participants, let us at least become familiar with the record of their behavior: the data. Examine them from every angle. Analyze the sexes separately. Make up composite indexes. If a datum suggests a new hypothesis, try to find additional evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you do not like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something—anything—interesting." (p. 187)

> To the extent our research practices embody the spirit of Bem's comments, is the extent to which we are living in the Garden of Forking Paths and the extent to which our statistical claims are suspect.

In addition to the researcher flexibility described above that is common to nearly all forms of research, psychophysiology research includes many decisions that could create problematic flexibility. Some of these are outlined in Luck and Gaspelin (in press; including some suggestions for reducing bias and error) and will not be re-iterated in depth here. I highlight them here to illustrate the forking paths present in ERP/EEG studies. For example, researchers vary in how they choose the epochs for analysis. Some might choose the epoch by examining waveforms and visually comparing points on grand average waveforms (essentially making multiple comparisons prior to doing the analysis). A problem Luck (2014) termed the *problem of multiple implicit comparisons*. Epoch calculations may also be changed post-hoc after examining results or waveforms. For example, in ERN research, one may measure the ERN between 0 and 100 ms but extend the epoch to 125 ms after looking at the waveforms and noticing that some peaks extend beyond 100 ms.

Researchers must determine the number of trials included in the grand average for each participant. This includes decisions about what constitutes too few trials and whether data should be analyzed with and without excluded participants. ERP researchers using high-density electrode configurations must choose the number of sensors to include and the location of the sensors. Researchers may choose just one electrode, but the reason for choosing that electrode may vary—the electrode with the maximal amplitude of the component, the sensor other researchers have used, or the electrode that produces the cleanest results. High density electrode nets mean that researchers can average over multiple sensors and researchers can vary in how averages are chosen. Perhaps researchers make an average region of interest and adjust the region of interest if the results are not quite statistically significant. Finally, ERP researchers also have to choose the amount of baseline correction and this can vary across studies, often with no clear reason for why the baseline values were chosen.

Most of these decisions will likely be made in good faith and be well intentioned. Unfortunately, the consequences of researcher flexibility are problematic regardless of the intentions of the researchers. Specifically, the consequence of research flexibility is an unreliable literature with an overabundance of positive and exaggerated findings. Indeed, the number of positive findings in the published literature increased 22% from 1990 to 2007 (Fanelli, 2012). Of the 100 pre-registered, well-powered studies in the Replication Project, only 36% found statistically significant effects. Likewise, estimated effect sizes were reduced by half (Open Science Collaboration, 2015). Taken together these results are consistent with the idea that research practices, such as The Garden of Forking Paths, are not producing stable effects.

4. Power and precision

The strength of any scientific literature rests squarely on the reliability and quality of the evidence. Psychophysiology, as well as most areas of psychology, use *p*-values as the primary statistical source of evidence. The *p*-value is used to reduce the number of false positives by ruling out chance as an explanation of the observed effects. Presumably, if the α-level of the study was set to 0.05, then false-positives should occur only 5% of the time. But what if the *p*-value does not in fact tell us whether an effect is real and the rate of false positives is actually higher? Further, what if the combination of reliance on *p*-values and the incentives in academic culture combine to create a situation where the published literature exaggerates effects? This would be a problem for most scientific disciplines.

4.1. Positive predictive value

To illustrate the way the published literature can exaggerate effects, Ioannidis (2005) introduced the positive predictive value (PPV) of a study, which is the probability that rejecting the null hypothesis with a statistical significance test reflects a true effect. PPV is function of four quantities: (a) the Type I error rate (α); (b) power (1-β), where β is the Type II error rate; (c) the ratio of true effects to no effects in a field (R); and (d) bias (u), which is defined as "the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced" (Ioannidis, 2005, p. 0697). Bias is roughly identical to researcher flexibility discussed previously. The PPV is computed as (see Ioannidis, 2005, Table 2, p. 0697):

$$PPV = \frac{(1-\beta)R + u\beta R}{(1-\beta)R + \alpha + u(1-\alpha+\beta R)}$$

Fig. 2 shows the relationship between PPV, power, bias, R, and α. The lines for bias = 0 are included as a reference for "relatively clean" studies (e.g., pre-registration of methods, excellent measures). As bias increases, the PPV decreases, so much so that even with 80% power, 1-to-1 pre-study odds that the effect is real, and bias at 30%, the PPV does not exceed 0.75. As power increases, the PPV increases, highlighting the critical role of power beyond just establishing statistical significance. That is, when power is relatively high, we have more confidence that statistically significant effects represent true effects. Unfortunately, power is typically low.

Button et al. (2013) reviewed neuroscience meta-analyses published in 2011 and found that the average power for neuroscience studies was 21%. Button et al. reviewed meta-analyses in neuroscience generally and their results do not speak specifically to psychophysiology. In fact, I am unaware of any systematic review of sample sizes and power in psychophysiology. Nevertheless, given that many psychological studies have low power to detect anything but large effects (Button et al., 2013; Cohen, 1962), it seems likely that psychophysiology studies also are underpowered.

To provide some data about sample sizes and power in psychophysiology, I present two samples of studies. First, I present sample sizes from an ongoing meta-analysis comparing error-related negativity (ERN) in younger and older participants. Second, I present the sample sizes from the final issue in 2015 of *Psychophysiology* and the *International Journal of Psychophysiology*. I excluded one methodological study (Zoumpoulaki et al., 2015) and two correlational studies (Palagini et al., 2015; Tomfohr et al., 2015) because they were too unique to compare to the bulk of the studies.

A study does not have a single power value because the power is a function of the specific analysis; a single study can include within- and between-subject contrasts, omnibus and pairwise comparisons, and correlation and regression analyses. Given the variability of study design across the samples, I focused on the power that each study had for a two condition contrast. For studies with a between-subjects focus, the contrast was between two groups (e.g., old vs young in the ERN studies). For studies with a within-subjects focus, the contrast was between two conditions (i.e., two timepoints, two types of stimuli). Most studies included an omnibus test (e.g., an *F*-test for the main effect of condition), but the form of that test varied considerably. However, nearly all studies included follow-up tests comparing two conditions

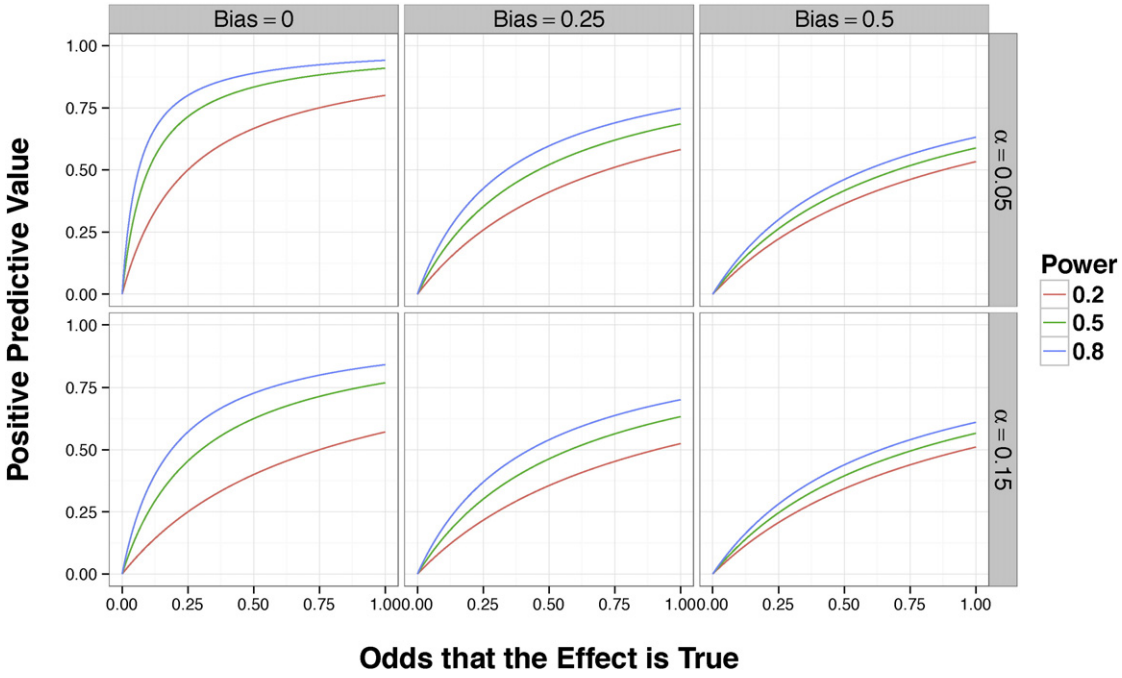

Fig. 2. Relationship between the positive predictive value of a study, power, bias, the prior odds that the effect is true, and Type I error rate (α).

in the form of post-hoc tests following an ANOVA. Furthermore, these condition specific contrasts are theoretically interesting contrasts (cf. Wampold et al., 1990).

Table 1 presents citation information and sample size for the ERN, *Psychophysiology*, and *International Journal of Psychophysiology* studies. For the ERN studies, the key contrast involves the contrasts between young and old participants and thus the key power to consider is the between-subjects contrast. The other studies are divided into within- and between-subject contrasts. Fig. 3 presents power curves for various sample sizes and effect sizes. For the ERN and between-subject curves, the sample size represents the number of participants per condition. For the within-subject curves, the sample size is the number of participants.

I included a dot and study number (see Table 1) so that each study can be located in Fig. 3. When interpreting Fig. 3, it is critical to remember that a study does not have a single power value. Instead, find a study, which has a specific sample size, and determine what power that sample would have to detect an effect of a given size. For example, consider the within-subject power curve. Study 31 (Marinovic et al., 2015) had a sample size of 15. A study of that size would have around 10% power to detect an effect of $d = 0.2$, around 45% power to detect and effect of $d = 0.5$, and around 90% power to detect an effect of $d = 0.8$. Study 29 (Kunecke et al., 2015) had a sample size of 48. Such a study would have over 80% power to detect effects >0.5, but still would be quite underpowered for small effects.

Looking over Fig. 3, it is apparent that most studies are well underpowered to detect all but large effects. Expected effect sizes will vary across substantive areas and meta-analyses can help narrow down an expected effect size. For example, a recent meta-analysis of the ERN and anxiety suggested that the average effect size comparing anxious participants to controls was $d = -0.36$ (Moser et al., 2016). Moderator analyses suggested that the effects for obsessive–compulsive symptoms were larger ($d = -0.64$) than anxiety symptoms ($d = -0.21$). Using the data from Moser et al. (2016), the average number of participants per condition for the overall effect ($d = -0.36$) was 20. Using the between-subject curves from Fig. 3, this suggests that power of the average study was between 30 and 35%. The average number of participants per condition for the obsessive–compulsive symptoms was 16.2, suggesting that the power of the average study was between 35% and 40%. Finally, the average number of participants per condition for the anxiety disorders symptoms was 22, suggesting that the power of the average study was about 10%. Moser et al. (2016) also showed that sex moderates the size of the effects, so power could be slightly different if the analysis changed. However, splitting the sample by sex only further reduces the sample size, which likely means power would be smaller.

An increase in α also reduces the PPV. What circumstances lead to increased α in psychophysiology research? Violation of the independence of observations assumption common to many statistical models, such as ANOVA or regression, leads to inflated Type I errors (Baldwin et al., 2005). These violations can occur in psychophysiology research that measure multiple members of a family. See Carrasco et al. (2013) for an example that used multiple members of a family but used statistical methods that control for non-independent observations.

Type I errors also increase when researchers perform multiple comparisons. The most common situations involving multiple comparisons are (a) post hoc tests following ANOVA or (b) running independent ANOVA or regression models on many different outcomes. Although most statistical texts cover methods for controlling the experiment-wise Type I error rate (e.g., Maxwell and Delaney, 2003), the focus is often on the use of appropriate follow-up tests for ANOVAs that maintain the experiment-wise Type I error rate. Unfortunately, psychophysiology studies often report analyses with a high number of uncorrected multiple comparisons, which can inflate the Type I error rate above the nominal level (Luck & Gaspelin, in press).

Cramer et al. (2015) showed that a multiway ANOVA, such as a 2×3 analysis, can have Type I error rates that can be as high as 14%, if the researchers do not have a specific hypothesis going into the analysis (see page 2). Absent a specific hypothesis, a 2×3 ANOVA has three tests of interest: two main effects and an interaction. Even if the null hypothesis is true, there can be a 14% chance that one of the three tests will be significant. Multiway ANOVAs with three factors can have Type I error rates near 30% and with four factors near 54%.

The notion of having a hypothesis versus not having a hypothesis stems from the distinction between exploratory and confirmatory research. In a confirmatory study, a researcher may use a 2×3 factorial experiment where theoretical predictions only involve the interaction. The Type I error rate for the interaction would still be set at 0.05

Table 1
Sample sizes for 51 psychophysiology studies.

Error related negativity and age study	Older age N	Younger age N
1. Band and Kok (2000)	13	17
2. Beste et al. (2009)	17	15
3. Capuana et al. (2012)	18	22
4. Endrass et al. (2012)	22	22
5. Eppinger and Kray (2011)	29	29
6. Eppinger et al. (2008)	18	18
7. Falkenstein et al. (2001)	11	11
8. Gehring and Knight (2000)	10	10
9. Herbert et al. (2011)	15	15
10. Hoffmann and Falkenstein (2011)	16	20
11. Kolev et al. (2005)	11	10
12. Mathalon et al. (2003)	10	10
13. Mathewson et al. (2005)	16	16
14. Nieuwenhuis et al. (2002)	13	16
15. Padilla (2005)	19	19
16. Pietschmann et al. (2011b)	18	15
17. Pietschmann et al. (2011a)	24	17
18. Pietschmann et al. (2008)	23	23
19. Schreiber et al. (2012)	16	16
20. Schreiber et al. (2011)	20	20
21. Staub et al. (2014)	16	15
22. Themanson et al. (2006)	32	34

Psychophysiology – within subjects study	N
23. Akyurek and van Asselt (2015) 1	18
24. Akyurek and van Asselt (2015) 2	19
25. Akyurek and van Asselt (2015) 3	31
26. Brown et al. (2015)	18
27. Hoenen et al. (2015)	30
28. Verkuil et al. (2015)	21
29. Kunecke et al. (2015)	48
30. Kuper et al. (2015)	20
31. Marinovic et al. (2015) 1	15
32. Marinovic et al. (2015) 2	24
33. Marinovic et al. (2015) 3	12
34. Marinovic et al. (2015) 4	12
35. Rodríguez-Herreros et al. (2015)	20
36. Sege et al. (2015)	30

Psychophysiology – between subjects study	N1	N2
37. Bradford et al. (2015)	48	48[a]
38. Verkuil et al. (2015)	19	41
39. Yoon et al. (2015)	74	77[b]

International Journal of Psychophysiology – within subjects study	N
40. Amico et al. (2015)	18
41. Choi et al. (2015)	13
42. Cutmore et al. (2015)	16
43. Ervast et al. (2015)	12
44. Jodoin et al. (2015)	21
45. MacDonald et al. (2015)	16
46. Meng and Ma (2015)	18
47. Zhu et al. (2015)	16

International Journal of Psychophysiology – between subjects study	N1	N2
48. Ceklic and Bastien (2015)	39	29
49. Chuang et al. (2015)	10	8
50. Parks et al. (2015)	48	50
51. Ramírez et al. (2015)	21	24

Note. a. Averaged over the placebo and no-alcohol conditions; b. Average of the psychopathology conditions.

(or other chosen level). However, the Type I error rate for the main effects would require adjustment. In contrast, in an exploratory study, a researcher may use a 2 × 3 factorial experiment but not have specific predictions about the main effects and interaction or it could be that any of the effects could support a hypothesis (cf. Gelman and Loken, 2014). In that case, a correction, such as a Bonferroni correction, would need to be applied to all three effects to ensure the experiment-wise error rate remains at 0.05 (Cramer et al., 2015). These challenges become more pronounced as the number of effects increases due to a more complicated design or multiple outcomes.

For example, in the error monitoring literature, a number of studies have examined the role age plays in the response to errors and many of these studies use a multiway ANOVA. Band and Kok (2000) used mental-rotation task to investigate how age affects the processing of errors. Their ANOVA model for reaction time used a 2 (young vs. old) × 3 (response deadline) × 2 (45 vs. 135 rotation angle) × 9 (reaction time bin) design. This produces 14 effects (4 main effects, 6 two-way interactions, 3 three-way interactions, and 1 four-way interaction) and this does not include the other outcome measures examined (see pages 204 and 206). Although the authors describe predictions at the beginning of their study, they do not map these predictions onto the various effects or provide a rationale of which effects were critical tests of their theory and which are exploratory.

Luck and Gaspelin (in press) describe the general problems with multiple comparisons in ERP research as follows:

> Common approaches to the statistical analysis of ERP experiments tend to lead to very high familywise and experimentwise error rates. These error rates are elevated in many ERP analyses relative to behavioral analyses because each condition of an experiment that yields a single behavioral measurement may yield many ERP measurements because multiple different components may be measured, because the amplitude and the latency may be measured for each component, and because each component may be measured at multiple electrode sites.This is a very common issue in ERP studies, including our own. To take an extreme example, consider a paper published by Luck and Hillyard (1994a) in *Psychophysiology* that provided a detailed analysis of several different ERP components in a set of visual search tasks. The analysis of the first experiment alone involve four 5-factor ANOVAs, three 4-factor ANOVAs, and three 3-factor ANOVAs, for a total of 190 individual *p* values. The experimentwise error rate for an experiment with 190 *p* values is close to 100%, so it is a near certainty that at least one of the significant effects in the experiment was bogus. (p. 19–20)

Of course, these examples do not prove that there was a problem with Band and Kok, Luck and Hillyard, or any psychophysiology study. However, when many effects estimated over many outcomes, the Type I error rate may not remain at the 0.05 level. My point in providing these examples is to show that Type I error rates may be elevated in some designs, which reduces the PPV of the findings.

Endrass et al. (2012) provide an example of how elevated Type I error rates can be addressed. They examined the relationship between age and error processing using a multiway repeated measures ANOVA. The between-subjects factor was age (old vs young). They had two within-subjects factors: condition (accuracy vs speed) and response type (correct vs incorrect). To maintain the Type I error rate, they used a Bonferroni-correction for all *p*-values.

As a supplement to this paper, I have created an interactive website that illustrates how PPV is affected by power, bias, pre-study odds that the effect is true, and Type I error rate. The website is an interactive version of Fig. 2 and allows users to vary parameters (e.g., power) to learn about the parameters impact on the PPV. The website can be found at: http://shinyserver.byu.edu/ppv/.

4.2. Exaggeration of effects

The discussion of the PPV of a study illustrated that research findings based on statistical significance tests do not always translate to true findings, especially when power is low and researchers use methods and analysis choices that introduce bias. The PPV treats findings as binary: either a finding is true or it is false. Another problem when power is low is that statistically significant effects tend to exaggerate real effects,

leading to the problem where the published literature is full of effects that are too big.

The potential for publication bias has been known for a long time (Greenwald, 1975). Publication bias typically refers to the fact that studies with statistically significant results are more likely to be published than studies with null results (Greenwald, 1975). This type of publication bias is referred to as the "file-drawer" problem (Rosenthal, 1979) because researchers keep null findings in their file-drawer rather than submit them for publication.[2] Although this type of publication bias is important and has been shown to reduce estimates of effect size (Driessen et al., 2015), this type of publication bias is not my focus.

Instead, my focus is on publication bias that produces exaggerated effect sizes or even effect estimates in the wrong direction, even within a single study. These types of errors have been called Type M (magnitude) and Type S (sign) errors (Gelman and Carlin, 2014). Type S errors occur when the reported effect is in the wrong direction (e.g., positive correlation when it should be negative). Type M errors occur when the reported effect exaggerates the true effect by either over- or under-estimating the true effect.

Like publication bias that occurs because authors, reviewers, and editors prefer statistically significant findings over non-significant findings, Type M and Type S errors occur because scientists tend to focus primarily, and sometimes solely, on the statistical significance of effects when drawing conclusions. That is, an effect is determined to be real if it is statistically significant without consideration of other methodological factors that suggest that the effect is overestimated and imprecise.

For example, Ursu et al. (2003) used fMRI methods to study activity in the anterior cingulate cortex (ACC) following errors in 11 patients with obsessive–compulsive disorder (OCD). The authors report the correlation between ACC activity and clinician-rated severity of OCD symptoms as measured by the Yale-Brown Obsessive-Compulsive Scale. Among the 11 patients, the correlation was $r = 0.46$. Ursu et al. report that the one-tailed p-value was 0.08 (95% confidence interval is $-0.19, 0.83$).[3] Although the authors are careful to note that this correlation is not statistically significant—they referred to it as a trend[4]—they do report and interpret the coefficient as evidence of a relationship.

The null hypothesis for the correlation coefficient is that the population correlation is 0. A one-tailed p-value of 0.08 indicates that *if the null hypothesis is true*, the probability of observing a correlation of $r = 0.46$ or larger is 0.08. The p-value does not provide any indication of the precision or quality of the correlation.

A key question when examining a point-estimate of $r = 0.46$ is whether we believe that the estimate is a good approximate of the population correlation (i.e., the actual correlation between AAC activity and OCD symptoms in the population). This depends on the value of the population correlation, which is by definition unknown. What can be approximated, however, is the sampling distribution of the correlation coefficient, which represents the variability in the correlation estimates one expects from one sample to another given a specific sample size. When samples are small, such as $N = 11$, the sampling distribution is wide, indicating a lot variability from study to study in the correlation estimate. Consequently, estimates from small studies are noisy, irrespective of the p-value.

Fig. 4 presents 10,000 draws from the sampling distribution for three sample sizes: $N = 11$, $N = 50$, and $N = 150$. The sampling distributions are centered at a population correlation of $\rho = 0.1$ to represent a small population correlation. I chose a small population correlation to illustrate that small studies can produce big estimates that are judged significant, even if the population effect is small. However, the degree to which a study exaggerates an effect is largely a function of sample size (Gelman and Carlin, 2014) rather than the specific population value because small studies are noisy.

As seen in Fig. 3, estimates from studies using a sample size of 11 vary from study-to-study far more than studies using a sample size of 50 or 150, increasing the probability of an exaggerated effect. Indeed, 12.5% of correlations are larger than 0.46 when $N = 11$ whereas 0.3% are larger than 0.46 or when $N = 50$ and 0% are larger than 0.46 when $N = 150$. When this information is combined with the fact that the only way to reject the null hypothesis when $N = 11$ is to obtain a large correlation, then using p-values in combination with small studies (i.e., low statistical power) creates a situation where effects judged to be important are likely to be exaggerated (Button et al., 2013; Gelman and Carlin, 2014).

Given these concerns, Ioannidis (2008) suggested that researchers consider the following questions when evaluating research:

> For a new proposed association, credibility and accuracy of its proposed effect varies depending on the case. One may ask the following questions: does the research community in this field adopt widely statistical significance or similar selection thresholds for claiming research findings? Did the discovery arise from a small study? Is there room for large flexibility in the analyses? Are we unprotected from selective reporting (e.g., was the protocol not fully available upfront)? Are there people or organizations interested in finding and promoting specific "positive" results? Finally, are the counteracting forces that would deflate effects minimal? (p. 644)

Clearly, these questions provide a more robust analysis of results than whether the effect was statistically significant. To be sure, many researchers and peer-reviewers consider alternative explanations and methodological factors that can influence results. However, considering Ioannidis' additional questions could help improve the accuracy of the psychophysiological literature.

A primary method for increasing the precision of estimates is to increase power (Button et al., 2013; Gelman and Carlin, 2014). Researchers have called for increasing power in psychological studies for a long time (Cohen, 1962), but it has not affected the practice of psychological research. Power is nearly always discussed in the context of rejecting the null hypothesis. Thus, if researchers obtain a statistically significant effect in their study, they, reviewers, and editors have shown that they are not likely to worry about power and the probability of exaggeration of effects. If researchers do not obtain a statistically significant effect, they, reviewers, and editors may raise lack of power as an explanation. As we have seen, the p-value does not provide evidence of whether an effect is noisy or exaggerated. Consequently, the focus should be on whether the effect is a good estimate, statistically significant or not.

5. Suggestions

So what do we do about these problems? A main theme of recent research and thinking about replication and rigor in research is how to change the fundamental practices of research and the incentives around research (Ioannidis, 2014; Simmons et al., 2011). Although individual scientists can make small changes, improvement will not come without major changes to the culture of science. As noted above, careers are built on papers and grants. Anything that threatens the process of publication and funding, will be seen as a threat to one's career, which in turn makes change difficult. However, if the culture changes, even bit-by-bit, then

[2] Or they submit their null findings but the paper is rejected for the lack of significant findings or contribution. They eventually give up and do not submit further.

[3] The fact that this is a one-tailed p-value may provide additional evidence of the power of p-values. Few papers consistently use one-tailed values. For example, Ursu et al. (2003) alternate between one- and two-tailed p-values in their various tests. One-tailed values sometimes appear when the two-tailed value would not quite be significant. That is, one-tailed values are often between 0.025 and 0.05, which would be consistent with two-tailed values between 0.05 and 0.1 (see McNally et al., 2004, for another psychophysiology example of one-tailed values in the 0.025 and 0.05 range). Although researchers may have chosen to use one-tailed tests prior to conducting the tests, it is also possible that the one-tailed tests allow the researchers to present the results as significant when the two-tailed tests would not (see the discussion of The Garden of Forking Paths above).

[4] For a humorous discussion of ways scientists deal with p-value above but near 0.05, see https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/.

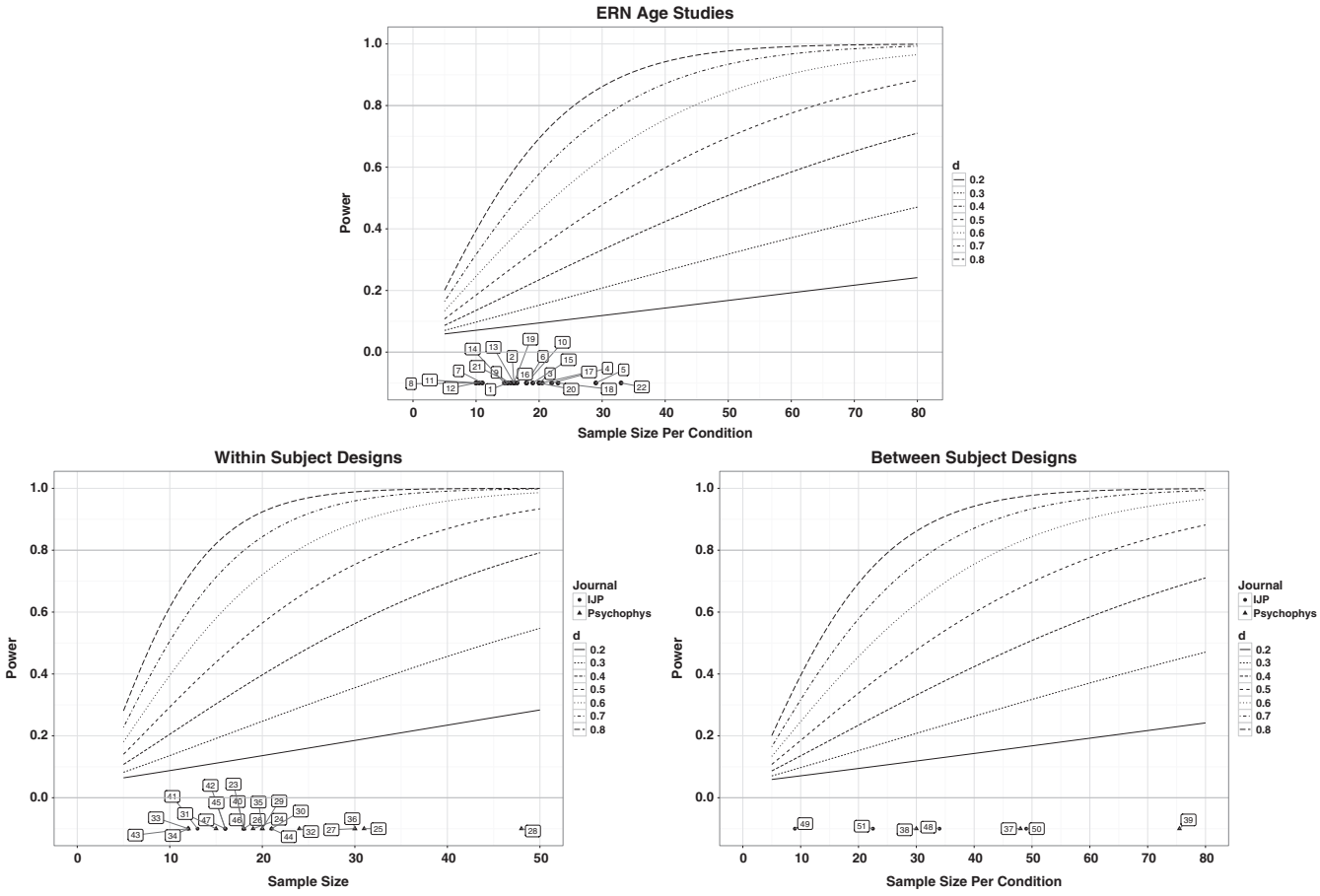

Fig. 3. 10,000 draws from the sampling distribution of a correlation coefficient for three sample sizes. The solid vertical line represents the population correlation and the dashed vertical line is the estimate from Ursu et al. (2003).

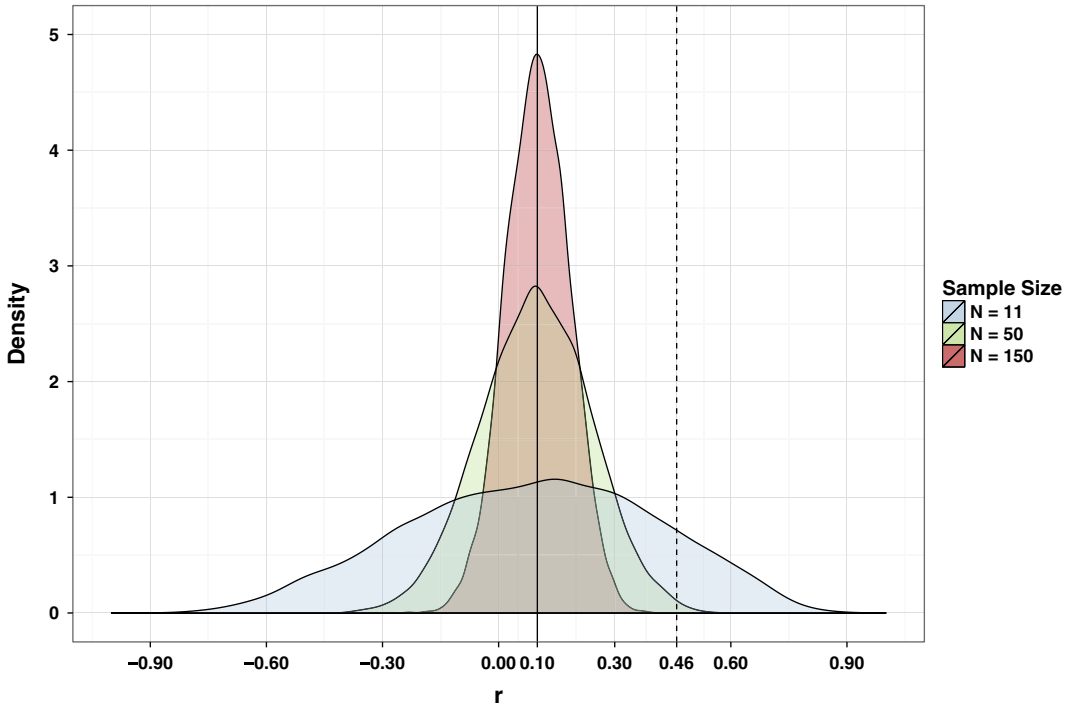

Fig. 4. Power as a function of sample size and effect size. The points represent the sample size for a given study. Study numbers can be found in Table 1. IJP = *International Journal of Psychophysiology*; Psychophys = *Psychophysiology*.

the process by which papers and grants are evaluated will also change—researchers will adapt. Below I review five suggestions from the literature for improving the quality of research. These suggestions are not groundbreaking. Instead they focus on the fundamentals of good scientific practice and address authors, reviewers, and editors. Although I wish there were two or three novel changes that would radically change the science for the better, I do not think such changes exist. However, a focus on these fundamentals should increase the probability of a solid evidence base in psychophysiology.

5.1. Increased power through collaboration

Although it is easy to recommend that researchers increase power in their studies, it is hard to actually increase power. A primary method for increasing power is to collaborate. Multiple groups working on the same problem can double or triple the number of participants for a particular study, without costing more money for any given research group. For example, several research groups study error processing using EEG. For the most part, these groups function independently, each producing relatively small studies—see Table 1 and Moser et al. (2016). It may be financially challenging for a given group to increase sample sizes. However, pooling data across labs could easily strengthen the power of a study.

I suspect the primary hurdle to increased collaboration is determining who will receive the credit. That is, if two independent groups can each publish a first-authored paper working on the same project, would it make sense to collaborate and only produce one paper? From a scientific perspective, the answer is an unequivocal yes. One only needs to review Figs. 2 and 4 to remember that under-powered studies create excess false-positives. Furthermore, if hiring, promotion, and grant committees provide more weight to reproducibility and precision of findings, rather than publication of findings (Ioannidis, 2014), scientists will gladly trade a first-authored, under-powered publication for a co-authored, well-powered publication.

5.2. Improved statistical and methodological training

Research is a skill. Like any skill, expertise in methodology develops over time, but researchers can get stuck in a rut of one or two methods and never develop their skills nor develop perspective that comes from using new and alternative methods (e.g., only using ANOVA, when mixed-models or even structural equation modeling may better address their question). Thus, researchers may become proficient at using their preferred method in ways that will lead to publication but may not be aware of important nuances of their methods (see previous section on multiplicity in factorial ANOVA).

Improved statistical skills begin with undergraduate and graduate training. A survey of quantitative training in psychology doctoral programs, documented substantial deficiencies in statistical and methodological training during graduate school (Aiken et al., 2008). For example, only 36% of all programs provided in-depth training in power analysis, despite the fact that power analysis and sample size calculation is essential to replicable, rigorous research. Furthermore, training in advanced methods, such as basic and advanced psychometrics, structural equation modeling, and multilevel modeling were not widely available in training programs.

Scientists need better training in conducting power analyses and understanding the role sample size plays in the precision and PPV of their results (e.g., Ioannidis, 2014). I suspect that many researchers do not conduct formal power analyses prior to data collection. If researchers are performing power analyses, they are not reporting them (Larson & Carbine, under review). In my role as a methodological consultant, I often see colleagues determine the sample size needed for their study by examining what sample sizes have led to published papers. Thus, if published papers have 20 participants, they will chose 20 participants. Unfortunately, whether a sample size has been published or is publishable, is not a good measure of sufficient power given that many studies are underpowered (Button et al., 2013) as well as the fact that researcher degrees of freedom and other publication biases can lead to significant effects (Luck & Gaspelin, in press).

5.3. Pre-registration of studies

The Garden of Forking Paths (see Fig. 1) illustrates the many decisions made during a study (e.g., Ioannidis, 2014; Open Science Collaboration, 2015; Simmons et al., 2011; Wagenmakers et al., 2012). To mitigate the consequences of these decisions and to increase transparency in the conduct and reporting of research, many have called for pre-registration of studies. Pre-registration involves stating hypotheses, sample sizes, measures, manipulations, and analysis plans prior to conducting a study. Doing so reduces the number of decisions researchers must make once they have seen the data and can increase our confidence in results. The Open Science Framework (http://www.osf.io) is free, flexible software for registering studies and it integrates with major software services that are key to many researchers' workflows (e.g., Dropbox, Google Drive).

5.4. Improved reporting standards

Along with pre-registrations, researchers can also increase the transparency of reporting in their studies. Given the habits of researchers and cultural norms about what is required to "get published," it is likely incumbent upon editors and reviewers to require better reporting in all studies in order for change to occur on a broad scale. To facilitate this change, the Open Science Framework has suggested a reviewer statement to be included in manuscript reviews (https://osf.io/hadz3/wiki/home/):

> I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure the statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open Science [see http://osf.io/hadz3]. I include it in every review.

Some journals, such as *Psychological Science*, have added disclosure statements to their manuscript submission process. These statements require that authors disclose that they have reported all measures analyzed, all conditions, methods for sample size calculations, and criteria for data exclusions (http://www.psychologicalscience.org).

5.5. Shifting incentives

As noted above, researchers are rewarded with jobs, promotion, prestige, and grant money for voluminous publishing. Given the potential that many findings are not true or are exaggerated, such rewards may not be useful. Consequently, some have argued that the scientific reward be shifted away from voluminous publishing toward reproducible research and carefully conducted replication research, among other things (Ioannidis, 2014; Nosek et al., 2012). In their paper, "Let's publish *fewer* papers", Nelson et al. (2012) argue that science would be better off if researchers were limited to publishing a single paper a year.

> As a thought experiment, consider a different utopia. In this one, researchers are allowed to publish only one paper per year. Publication quantity is no longer a relevant dimension. This system incentivizes researchers to demonstrate that an effect is robust and generalizable, and hence true and important. Rather than the community of researchers being forced to wade through a mountain of papers to discern, with extreme difficulty, the true ones from the false ones, it is the researcher herself who chooses among all of the effects she would have attempted to publish in order to focus on the one that

she can obtain most reliably. (p. 292)

Better incentives for high-quality peer-review, both pre- and post-publication peer-review are needed (Ioannidis, 2014; Nosek et al., 2012). Reviewing is a largely thankless job. Given the massive increase in the number of journals across psychology, along with the ease of electronic submission of manuscripts, the opportunities for thankless work abound. With the exception of listing oneself as a reviewer or editorial board member on one's vita, there is little professional incentive for thorough, thoughtful reviews. To be sure, many reviewers are at least partially motivated by the desire for a reliable, robust literature. However, given the demands on researchers' time and the huge number of manuscripts, it is not surprising that some reviews are not particularly thoughtful or rigorous.

Peer-review need not end with publication. Post-publication peer-review is important and can be a useful mechanism for uncovering problems in a research area (e.g., Bhattacharjee, 2013; Bohannon, 2015). However, post-publication peer-review is not always valued. For example, researchers who have focused on critically examining the reproducibility of psychological studies have been labeled "bullies" (M. N. Meyer and Chabris, 2014) or engaging in "negative psychology" (Coan, 2014). Those focused on improving methods in psychology may sometimes get personal and attack researchers rather than the research. However, my impression is that most critics of psychology are aiming to improve the quality of ideas in the discipline. Further, better incentives for post-publication peer review, in the form of commentaries and replication studies, will help make this critical part of the scientific process normative (Open Science Collaboration, 2015).

6. Conclusion

Are psychology and related fields really in a crisis? I do not know. What is clear is that our current research practices are not producing reliable, reproducible research at the rate we would like. The nature of incentives in academics, combined with research practices and norms that reduce the probability of strong findings (e.g., underpowered studies), suggests a cultural change is needed to make lasting changes. The first step in a cultural change is awareness. I believe the evidence is clear that a change is needed—what is left, is the hard work to change.

References

Aiken, L.S., West, S.G., Millsap, R.E., 2008. Doctoral training in statistics, measurement, and methodology in psychology: replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. Am. Psychol. 63, 32–50. http://dx.doi.org/10.1037/0003-066X.63.1.32.

Akyurek, E.G., van Asselt, E.M., 2015. Spatial attention facilitates assembly of the briefest percepts: electrophysiological evidence from color fusion. Psychophysiology 52, 1646–1663. http://dx.doi.org/10.1111/psyp.12523.

Amico, F., Ambrosini, E., Guillem, F., Mento, G., Power, D., Pergola, G., Vallesi, A., 2015. The virtual tray of objects task as a novel method to electrophysiologically measure visuo-spatial recognition memory. Int. J. Psychophysiol. 98, 477–489. http://dx.doi.org/10.1016/j.ijpsycho.2015.10.006.

Anderson, C.J., Bahník, Š., Barnett-Cowan, M., Bosco, F.A., Chandler, J., Chartier, C.R., ... Zuni, K., 2016. Response to comment on "estimating the reproducibility of psychological science". Science 351, 1037. http://dx.doi.org/10.1126/science.aad9163.

Baldwin, S.A., Larson, M.J., Clayson, P.E., 2015. The dependability of electrophysiological measurements of performance monitoring in a clinical sample: a generalizability and decision analysis of the ERN and Pe. Psychophysiology 52, 790–800. http://dx.doi.org/10.1111/psyp.12401.

Baldwin, S.A., Murray, D.M., Shadish, W.R., 2005. Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. J. Consult. Clin. Psychol. 73, 924–935. http://dx.doi.org/10.1037/0022-006X.73.5.924.

Band, G.P., Kok, A., 2000. Age effects on response monitoring in a mental-rotation task. Biol. Psychol. 51, 201–221.

Bem, D.J., 2004. Writing the Empirical Journal Article. In: Darley, J.M., Zanna, M.P., Roediger III, H.L. (Eds.), The Compleat Academic: A Career Guide, second ed. American Psychological Association, Washington, DC, pp. 185–219.

Bem, D.J., 2011. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. J. Pers. Soc. Psychol. 100, 407–425. http://dx.doi.org/10.1037/a0021524.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155. http://dx.doi.org/10.1111/j.1749-6632.2010.05446.x.

Beste, C., Willemssen, R., Saft, C., Falkenstein, M., 2009. Error processing in normal aging and in basal ganglia disorders. Neuroscience 159, 143–149. http://dx.doi.org/10.1016/j.neuroscience.2008.12.030.

Bhattacharjee, Y., 2013. The mind of a con man. (Retrieved April 11, 2016, from) http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?pagewanted=all&_r=0.

Bohannon, J., 2015. Science retracts gay marriage paper without agreement of lead author LaCour. (from) http://news.sciencemag.org/policy/2015/05/science-retracts-gay-marriage-paper-without-lead-author-s-consent.

Bradford, D.E., Starr, M.J., Shackman, A.J., Curtin, J.J., 2015. Empirically based comparisons of the reliability and validity of common quantification approaches for eyeblink startle potentiation in humans. Psychophysiology 52, 1669–1681. http://dx.doi.org/10.1111/psyp.12545.

Brown, S.B., van der Wee, N.J., van Noorden, M.S., Giltay, E.J., Nieuwenhuis, S., 2015. Noradrenergic and cholinergic modulation of late ERP responses to deviant stimuli. Psychophysiology 52, 1620–1631. http://dx.doi.org/10.1111/psyp.12544.

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. Nat. Rev. Neurosci. 14, 365–376. http://dx.doi.org/10.1038/nrn3475.

Capuana, L.J., Dywan, J., Tays, W.J., Segalowitz, S.J., 2012. Cardiac workload and inhibitory control in younger and older adults. Biol. Psychol. 90, 60–70. http://dx.doi.org/10.1016/j.biopsycho.2012.02.018.

Carrasco, M., Harbin, S.M., Nienhuis, J.K., Fitzgerald, K.D., Gehring, W.J., Hanna, G.L., 2013. Increased error-related brain activity in youth with obsessive–compulsive disorder and unaffected siblings. Depress. Anxiety 30, 39–46. http://dx.doi.org/10.1002/da.22035.

Ceklic, T., Bastien, C.H., 2015. Information processing during NREM sleep and sleep quality in insomnia. Int. J. Psychophysiol. 98, 460–469. http://dx.doi.org/10.1016/j.ijpsycho.2015.10.003.

Choi, D., Ota, S., Watanuki, S., 2015. Does cigarette smoking relieve stress? Evidence from the event-related potential (ERP). Int. J. Psychophysiol. 98, 470–476. http://dx.doi.org/10.1016/j.ijpsycho.2015.10.005.

Chuang, L.-Y., Huang, C.-J., Hung, T.-M., 2015. Effects of attentional training on visual attention to emotional stimuli in archers: a preliminary investigation. Int. J. Psychophysiol. 98, 448–454. http://dx.doi.org/10.1016/j.ijpsycho.2015.09.001.

Coan, J., 2014. Negative psychology — medium. (Retrieved April 11, 2016, from https://medium.com/@jimcoan/negative-psychology-f66795952859, May 30).

Cohen, J., 1962. The Statistical Power of Abnormal-Social Psychological Research: A Review. J. Abnorm. Soc. Psychol.

Cohen, J., 1994. The earth is round ($p < 0.05$). Am. Psychol. 49, 997–1003.

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. Psychon. Bull. Rev. 1–8 http://dx.doi.org/10.3758/s13423-015-0913-5.

Cutmore, T.R.H., Halford, G.S., Wang, Y., Ramm, B.J., Spokes, T., Shum, D.H.K., 2015. Neural correlates of deductive reasoning: an ERP study with the Wason selection task. Int. J. Psychophysiol. 98, 381–388. http://dx.doi.org/10.1016/j.ijpsycho.2015.07.004.

Driessen, E., Hollon, S.D., Bockting, C.L.H., Cuijpers, P., Turner, E.H., 2015. Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of US National Institutes of Health-funded trials. PLoS One 10, e0137864. http://dx.doi.org/10.1371/journal.pone.0137864.

Endrass, T., Schreiber, M., Kathmann, N., 2012. Speeding up older adults: age-effects on error processing in speed and accuracy conditions. Biol. Psychol. 89, 426–432. http://dx.doi.org/10.1016/j.biopsycho.2011.12.005.

Eppinger, B., Kray, J., 2011. To choose or to avoid: age differences in learning from positive and negative feedback. J. Cogn. Neurosci. 23, 41–52. http://dx.doi.org/10.1162/jocn.2009.21364.

Eppinger, B., Kray, J., Mock, B., Mecklinger, A., 2008. Better or worse than expected? Aging, learning, and the ERN. Neuropsychologia 46, 521–539. http://dx.doi.org/10.1016/j.neuropsychologia.2007.09.001.

Ervast, L., Hämäläinen, J.A., Zachau, S., Lohvansuu, K., Heinänen, K., Veijola, M., ... Leppänen, P.H.T., 2015. Event-related brain potentials to change in the frequency and temporal structure of sounds in typically developing 5–6-year-old children. Int. J. Psychophysiol. 98, 413–425. http://dx.doi.org/10.1016/j.ijpsycho.2015.08.007.

Falkenstein, M., Hoormann, J., Hohnsbein, J., 2001. Changes of error-related ERPs with age. Exp. Brain Res. 138, 258–262. http://dx.doi.org/10.1007/s002210100712.

Fanelli, D., 2012. Negative results are disappearing from most disciplines and countries. Scientometrics 90, 891–904. http://dx.doi.org/10.1007/s11192-011-0494-7.

Foti, D., Kotov, R., Hajcak, G., 2013. Psychometric considerations in using error-related brain activity as a biomarker in psychotic disorders. J. Abnorm. Psychol. 122, 520–531. http://dx.doi.org/10.1037/a0032618.

Gehring, W.J., Knight, R.T., 2000. Prefrontal-cingulate interactions in action monitoring. Nat. Neurosci. 3, 516–520. http://dx.doi.org/10.1038/74899.

Gelman, A., Carlin, J., 2014. Beyond power calculations assessing type S (sign) and type M (magnitude) errors. Perspect. Psychol. Sci. 9, 641–651. http://dx.doi.org/10.1177/1745691614551642.

Gelman, A., Loken, E., 2013. The garden of forking paths: why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. (Retrieved from) http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Gelman, A., Loken, E., 2014. The statistical crisis in science. Am. Sci. 102, 460–465.

Gilbert, D.T., King, G., Pettigrew, S., Wilson, T.D., 2016. Comment on "estimating the reproducibility of psychological science". Science 351, 1037. http://dx.doi.org/10.1126/science.aad7243.
Greenwald, A.G., 1975. Consequences of prejudice against the null hypothesis. Psychol. Bull. 82, 1–20. http://dx.doi.org/10.1037/h0076157.
Hajcak, G., Foti, D., 2008. Errors are aversive: defensive motivation and the error-related negativity. Psychol. Sci. 19, 103–108. http://dx.doi.org/10.1111/j.1467-9280.2008.02053.x.
Herbert, M., Eppinger, B., Kray, J., 2011. Younger but not older adults benefit from salient feedback during learning. Front. Psychol. 2, 171. http://dx.doi.org/10.3389/fpsyg.2011.00171.
Hoenen, M., Lubke, K.T., Pause, B.M., 2015. Somatosensory mu activity reflects imagined pain intensity of others. Psychophysiology 52, 1551–1558. http://dx.doi.org/10.1111/psyp.12522.
Hoffmann, S., Falkenstein, M., 2011. Aging and error processing: age related increase in the variability of the error-negativity is not accompanied by increase in response variability. PLoS One 6. http://dx.doi.org/10.1371/journal.pone.0017482 (e17482-17410).
Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2, e124. http://dx.doi.org/10.1371/journal.pmed.0020124.
Ioannidis, J.P.A., 2008. Why most discovered true associations are inflated. Epidemiology 19, 640–648. http://dx.doi.org/10.1097/EDE.0b013e31818131e7.
Ioannidis, J.P.A., 2012. Why science is not necessarily self-correcting. Perspect. Psychol. Sci. 7, 645–654. http://dx.doi.org/10.1177/1745691612464056.
Ioannidis, J.P.A., 2014. How to make more published research true. PLoS Med. 11, e1001747. http://dx.doi.org/10.1371/journal.pmed.1001747.
Ioannidis, J.P.A., Trikalinos, T.A., 2005. Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. J. Clin. Epidemiol. 58, 543–549. http://dx.doi.org/10.1016/j.jclinepi.2004.10.019.
Jodoin, V.D., Lespérance, P., Nguyen, D.K., Fournier-Gosselin, M.-P., Richer, F., 2015. Effects of vagus nerve stimulation on pupillary function. Int. J. Psychophysiol. 98, 455–459. http://dx.doi.org/10.1016/j.ijpsycho.2015.10.001.
Kolev, V., Falkenstein, M., Yordanova, J., 2005. Aging and error processing. J. Psychophysiol. 19, 289–297. http://dx.doi.org/10.1027/0269-8803.19.4.289.
Kunecke, J., Sommer, W., Schacht, A., Palazova, M., 2015. Embodied simulation of emotional valence: facial muscle responses to abstract and concrete words. Psychophysiology 52, 1590–1598. http://dx.doi.org/10.1111/psyp.12555.
Kuper, K., Liesefeld, A.M., Zimmer, H.D., 2015. ERP evidence for hemispheric asymmetries in abstract but not exemplar-specific repetition priming. Psychophysiology 52, 1610–1619. http://dx.doi.org/10.1111/psyp.12542.
Larson, M.J., Carbine, K.A., 2016. Sample Size Calculations in Human Electrophysiology (EEG and ERP) Studies: A Systematic Review and Recommendations for Increased Rigor (under review).
Larson, M.J., Baldwin, S.A., Good, D.A., Fair, J.E., 2010. Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): the role of number of trials. Psychophysiology 47, 1167–1171. http://dx.doi.org/10.1111/j.1469-8986.2010.01022.x.
Latour, B., Woolgar, S., 1986. Laboratory Life: The Construction of Scientific Facts. second ed. Princeton University Press, Princenton, NJ.
Lord, C.G., 2004. A Guide to PhD Graduate School: How they Keep Score in the Big Leagues. In: Darley, J.M., Zanna, M.P., Roediger III, H.L. (Eds.), The Compleat Academic: A Career Guide, second ed. American Psychological Association, Washington, DC, pp. 3–16.
Luck, S.J., 2014. An Introduction to the Event-Related Potential Technique. second ed. MIT Press, Cambridge, MA.
Luck, S. J., & Gaspelin, N. (2016). How to get statistically significant effects in any ERP experiment (and why you shouldn't). Psychophysiology, (in press).
MacDonald, B., Barry, R.J., Bonfield, R.C., 2015. Trials and intensity effects in single-trial ERP components and autonomic responses in a dishabituation paradigm with very long ISIs. Int. J. Psychophysiol. 98, 394–412. http://dx.doi.org/10.1016/j.ijpsycho.2015.08.002.
Marinovic, W., Milford, M., Carroll, T., Riek, S., 2015. The facilitation of motor actions by acoustic and electric stimulation. Psychophysiology 52, 1698–1710. http://dx.doi.org/10.1111/psyp.12540.
Mathalon, D.H., Bennett, A., Askari, N., Gray, E.M., Rosenbloom, M.J., Ford, J.M., 2003. Response-monitoring dysfunction in aging and Alzheimer's disease: an event-related potential study. Neurobiol. Aging 24, 675–685. http://dx.doi.org/10.1016/S0197-4580(02)00154-9.
Mathewson, K.J., Dywan, J., Segalowitz, S.J., 2005. Brain bases of error-related ERPs as influenced by age and task. Biol. Psychol. 70, 88–104. http://dx.doi.org/10.1016/j.biopsycho.2004.12.005.
Maxwell, S.E., Delaney, H.D., 2003. Designing Experiments and Analyzing Data: A Model Comparison Perspective. second ed. Taylor & Francis, New York.
McNally, R.J., Lasko, N.B., Clancy, S.A., Macklin, M.L., Pitman, R.K., Orr, S.P., 2004. Psychophysiological responding during script-driven imagery in people reporting abduction by space aliens. Psychol. Sci. 15, 493–497. http://dx.doi.org/10.1111/j.0956-7976.2004.00707.x.
Meng, L., Ma, Q., 2015. Live as we choose: the role of autonomy support in facilitating intrinsic motivation. Int. J. Psychophysiol. 98, 441–447. http://dx.doi.org/10.1016/j.ijpsycho.2015.08.009.
Meyer, M.N., Chabris, C., 2014. Why Psychologists' Food Fight Matters. (Retrieved April 11, 2016, from http://www.slate.com/articles/health_and_science/science/2014/07/replication_controversy_in_psychology_bullying_file_drawer_effect_blog_posts.html, Jul 31).
Meyer, A., Riesel, A., Proudfit, G.H., 2013. Reliability of the ERN across multiple tasks as a function of increasing errors. Psychophysiology 50, 1220–1225. http://dx.doi.org/10.1111/psyp.12132.
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. http://dx.doi.org/10.1111/psyp.12509.
Moser, J.S., Moran, T.P., Schroder, H.S., Donnellan, M.B., Yeung, N., 2014. The case for compensatory processes in the relationship between anxiety and error monitoring: a reply to Proudfit, Inzlicht, and Mennin. Front. Hum. Neurosci. 8, 64. http://dx.doi.org/10.3389/fnhum.2014.00064.
Nelson, L.D., Simmons, J.P., Simonsohn, U., 2012. Let's publish fewer papers. Psychol. Inq. 23, 291–293.
Nieuwenhuis, S., Ridderinkhof, K.R., Talsma, D., Coles, M.G.H., Holroyd, C.B., Kok, A., van der Molen, M.W., 2002. A computational account of altered error processing in older age: dopamine and the error-related negativity. Cogn. Affect. Behav. Neurosci. 2, 19–36.
Nosek, B.A., Spies, J.R., Motyl, M., 2012. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. Perspect. Psychol. Sci. 7, 615–631. http://dx.doi.org/10.1177/1745691612459058.
Olvet, D.M., Hajcak, G., 2009. The stability of error-related brain activity with increasing trials. Psychophysiology 46, 957–961. http://dx.doi.org/10.1111/j.1469-8986.2009.00848.x.
Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349. http://dx.doi.org/10.1126/science.aac4716 (aac4716-aac4716).
Padilla, M. L. (2005). *The role of prefrontal cortex in stimulus guided behavior: An EEG study of younger, older, and frontal brain damaged adults.* (Unpublished doctoral dissertation), University of California, Berkeley, Berkeley, CA.
Pagano, R.R., 2013. Understanding Statistics in the Behavioral Sciences. 10th ed. Wadsworth, Belmont, CA.
Palagini, L., Ragno, G., Caccavale, L., Gronchi, A., Terzaghi, M., Mauri, M., ... Manni, R., 2015. Italian validation of the sleep condition indicator: a clinical screening tool to evaluate insomnia disorder according to DSM-5 criteria. Int. J. Psychophysiol. 98, 435–440. http://dx.doi.org/10.1016/j.ijpsycho.2015.08.008.
Parks, A.C., Moore, R.D., Wu, C.-T., Broglio, S.P., Covassin, T., Hillman, C.H., Pontifex, M.B., 2015. The association between a history of concussion and variability in behavioral and neuroelectric indices of cognition. Int. J. Psychophysiol. 98, 426–434. http://dx.doi.org/10.1016/j.ijpsycho.2015.08.006.
Pashler, H., Harris, C.R., 2012. Is the replicability crisis overblown? Three arguments examined. Perspect. Psychol. Sci. 7, 531–536. http://dx.doi.org/10.1177/1745691612463401.
Pietschmann, M., Endrass, T., Kathmann, N., 2011a. Age-related alterations in performance monitoring during and after learning. Neurobiol. Aging 32, 1320–1330. http://dx.doi.org/10.1016/j.neurobiolaging.2009.07.016.
Pietschmann, M., Endrass, T., Czerwon, B., Kathmann, N., 2011b. Aging, probabilistic learning and performance monitoring. Biol. Psychol. 86, 74–82. http://dx.doi.org/10.1016/j.biopsycho.2010.10.009.
Pietschmann, M., Simon, K., Endrass, T., Kathmann, N., 2008. Changes of performance monitoring with learning in older and younger adults. Psychophysiology 45, 559–568. http://dx.doi.org/10.1111/j.1469-8986.2008.00651.x.
Pontifex, M.B., Scudder, M.R., Brown, M.L., O'Leary, K.C., Wu, C.-T., Themanson, J.R., Hillman, C.H., 2010. On the number of trials necessary for stabilization of error-related brain activity across the life span. Psychophysiology 47, 767–773. http://dx.doi.org/10.1111/j.1469-8986.2010.00974.x.
Ramírez, E., Ortega, A.R., Del Paso, G.A.R., 2015. Anxiety, attention, and decision making: the moderating role of heart rate variability. Int. J. Psychophysiol. 98, 490–496. http://dx.doi.org/10.1016/j.ijpsycho.2015.10.007.
Riesel, A., Weinberg, A., Moran, T., Hajcak, G., 2013. Time course of error-potentiated startle and its relationship to error-related brain activity. J. Psychophysiol. 27, 51–59.
Rodríguez-Herreros, B., Rodríguez-Fornells, A., López-Moliner, J., 2015. The neural correlates of motion-induced shifts in reaching. Psychophysiology 52, 1577–1589. http://dx.doi.org/10.1111/psyp.12519.
Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychol. Bull. 86, 638–641. http://dx.doi.org/10.1037/0033-2909.86.3.638.
Schreiber, M., Endrass, T., Weigand, A., Kathmann, N., 2012. Age effects on adjustments of performance monitoring to task difficulty. J. Psychophysiol. 26, 145–153. http://dx.doi.org/10.1027/0269-8803/a000077.
Schreiber, M., Pietschmann, M., Kathmann, N., Endrass, T., 2011. ERP correlates of performance monitoring in elderly. Brain Cogn. 76, 131–139. http://dx.doi.org/10.1016/j.bandc.2011.02.003.
Schwartz, S.J., Lilienfeld, S.O., Meca, A., Sauvigné, K.C., 2016. The role of neuroscience within psychology: a call for inclusiveness over exclusiveness. Am. Psychol. 71, 52–70. http://dx.doi.org/10.1037/a0039678.
Sege, C.T., Bradley, M.M., Lang, P.J., 2015. Prediction and perception: defensive startle modulation. Psychophysiology 52, 1664–1668. http://dx.doi.org/10.1111/psyp.12546.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359–1366. http://dx.doi.org/10.1177/0956797611417632.
Staub, B., Doignon-Camus, N., Bacon, É., Bonnefond, A., 2014. Age-related differences in the recruitment of proactive and reactive control in a situation of sustained attention. Biol. Psychol. 103, 38–47. http://dx.doi.org/10.1016/j.biopsycho.2014.08.007.
Themanson, J.R., Hillman, C.H., Curtin, J.J., 2006. Age and physical activity influences on action monitoring during task switching. Neurobiol. Aging 27, 1335–1345. http://dx.doi.org/10.1016/j.neurobiolaging.2005.07.002.
Tomfohr, L.M., Edwards, K.M., Madsen, J.W., Mills, P.J., 2015. Social support moderates the relationship between sleep and inflammation in a population at high risk for developing cardiovascular disease. Psychophysiology 52, 1689–1697. http://dx.doi.org/10.1111/psyp.12549.
Ursu, S., Stenger, V.A., Shear, M.K., Jones, M.R., Carter, C.S., 2003. Overactive action monitoring in obsessive–compulsive disorder: evidence from functional magnetic

resonance imaging. Psychol. Sci. 14, 347–353. http://dx.doi.org/10.1111/1467-9280.24411.
Verkuil, B., Brosschot, J.F., Marques, A.H., Kampschroer, K., Sternberg, E.M., Thayer, J.F., 2015. Gender differences in the impact of daily sadness on 24-h heart rate variability. Psychophysiology 52, 1682–1688. http://dx.doi.org/10.1111/psyp.12541.
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H.L.J., 2011. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). J. Pers. Soc. Psychol. 100, 426–432. http://dx.doi.org/10.1037/a0022790.
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H.L.J., Kievit, R.A., 2012. An agenda for purely confirmatory research. Perspect. Psychol. Sci. 7, 632–638. http://dx.doi.org/10.1177/1745691612463078.
Wampold, B.E., Davis, B., Good, R.H., 1990. Hypothesis validity of clinical research. J. Consult. Clin. Psychol. 58, 360–367.
Yoon, H.H., Malone, S.M., Iacono, W.G., 2015. Longitudinal stability and predictive utility of the visual P3 response in adults with externalizing psychopathology. Psychophysiology 52, 1632–1645. http://dx.doi.org/10.1111/psyp.12548.
Zhu, X., Gu, R., Wu, H., Luo, Y., 2015. Self-reflection modulates the outcome evaluation process: evidence from an ERP study. Int. J. Psychophysiol. 98, 389–393. http://dx.doi.org/10.1016/j.ijpsycho.2015.08.001.
Zoumpoulaki, A., Alsufyani, A., Filetti, M., Brammer, M., Bowman, H., 2015. Latency as a region contrast: measuring ERP latency differences with dynamic time warping. Psychophysiology 52, 1559–1576. http://dx.doi.org/10.1111/psyp.12521.