General Article

Advances in Methods and
Practices in Psychological Science
1–10
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245919858428
www.psychologicalscience.org/AMPPS

Should Psychology Journals Adopt Specialized Statistical Review?

Tom E. Hardwicke[1,2], **Michael C. Frank**[3], **Simine Vazire**[4], **and Steven N. Goodman**[1,5,6]

[1]Meta-Research Innovation Center at Stanford (METRICS), Stanford University; [2]Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Berlin Institute of Health and Charité – Universitätsmedizin Berlin, Berlin, Germany; [3]Department of Psychology, Stanford University; [4]Department of Psychology, University of California, Davis; [5]Department of Epidemiology, Stanford University School of Medicine; and [6]Department of Medicine, Stanford University School of Medicine

Abstract

Readers of peer-reviewed research may assume that the reported statistical analyses supporting scientific claims have been closely scrutinized and surpass a high-quality threshold. However, widespread misunderstanding and misuse of statistical concepts and methods suggests that suboptimal or erroneous statistical practice is routinely overlooked during peer review in psychology. Here, we explore whether psychology journals could ameliorate some of the field's statistical ailments by adopting specialized statistical review: a focused technical assessment, performed by statistical experts, that addresses the analysis and presentation of quantitative information and supplements regular peer review. We discuss evidence from a recent survey of journal editors suggesting that specialized statistical review may be unusual in psychology journals and is regarded by many editors as unnecessary. We contrast these views with those in the biomedical domain, where statistical review has been considered a partial preventive measure against the improper use of statistics since the late 1970s. We suggest that the current "credibility revolution" presents an opportune occasion for psychology journals to consider adopting specialized statistical review.

Keywords

peer review, statistical review, metaresearch, statistical methods, journal policy, open data, open materials

Received 12/1/18; Revision accepted 5/30/19

> After moving to a system of having a statistician present at every meeting, none of the editorial team could imagine moving back to a system where they were not present.
>
> —Richard Smith, former editor of the *British Medical Journal* (now *The BMJ*) (Smith, 2005, p. 2).

Scientific claims in psychology often rely on a scaffold of statistical analyses that support inductive inferences from samples of data (Rosnow & Rosenthal, 1989). The appropriate selection, implementation, reporting, and interpretation of these analyses is necessary for the validity of the associated claims (Cook & Campbell, 1979; García-Pérez, 2012). Readers of the peer-reviewed literature may assume that reported statistical analyses have been closely scrutinized for quality. But serious concerns about the credibility of psychological research have been raised (Baker, 2016; Pashler & Wagenmakers, 2012), and the misunderstanding and misuse of statistical methods has been implicated as an important cause (Button et al., 2013; Gigerenzer, 2018; Munafò et al., 2017; Simmons, Nelson, & Simonsohn, 2011).

In this article, we explore whether psychology journals could ameliorate some of the field's statistical ailments by

Corresponding Author:
Tom E. Hardwicke, Meta-Research Innovation Center Berlin, QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany
E-mail: tom.hardwicke@charite.de

adopting *specialized statistical review*: a focused technical assessment, performed by statistical experts, that addresses the analysis and presentation of quantitative information, supplementing regular peer review. In biomedicine, statistical review has been considered a partial preventive measure against the improper use of statistics since the late 1970s (Altman, 1982, 1994, 1998; Smith, 2005; Sox, 2009). In a recent survey, we found that 71 of 107 editors (66%) at leading biomedical journals reported that they routinely employed statistical review for 10% or more of submitted manuscripts, and 25 (23%) said they used statistical review for all manuscripts (Hardwicke & Goodman, 2019; also see George, 1985; Goodman, Altman, & George, 1998). By contrast, the survey responses from a sample of 39 psychology-journal editors, reported in this article, suggest that specialized statistical review is unusual in psychology journals and often regarded as unnecessary. We summarize evidence suggesting that statistical problems are commonplace in the published literature and discuss whether the apparent value of statistical review in biomedical journals could translate to psychology. We suggest that the current "credibility revolution" (Nelson, Simmons, & Simonsohn, 2018; Vazire, 2018) presents an opportune occasion for psychology-journal editors to consider adopting specialized statistical review.

Disclosures

Data, materials, and online resources

All data (https://osf.io/nquws/files/), survey materials (https://osf.io/tmah8/files/), and analysis scripts (https://osf.io/4zurk/files/) related to this study are publicly available on the Open Science Framework. To facilitate reproducibility, we wrote this manuscript by interleaving regular prose and analysis code, using knitr (Xie, 2018) and papaja (Aust & Barth, 2019), and have made the manuscript available in a software container (https://doi.org/10.24433/CO.8241121.v3) that recreates the computational environment in which the original analyses were performed. Detailed methods and results for the survey of psychology editors is provided in the Supplemental Material (available online at http://journals.sagepub.com/doi/suppl/10.1177/2515245919858428).

Reporting

The survey data reported here represent the subsample of psychology journals included in a broader survey of statistical-reviewing policies at biomedical journals. The findings for biomedical journals will be reported elsewhere (Hardwicke & Goodman, 2019), and the findings for psychology journals are reported here for the first time. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethical approval

This study was approved by the institutional review board of the Stanford University School of Medicine.

What Do Psychology-Journal Editors Think About Statistical Review? Results of a Survey

To gauge the current use of statistical review, we surveyed a sample of high-impact psychology journals (full methods and results are provided in the Supplemental Material available online). We received responses from editors (all but one an editor-in-chief) at 39 of 118 psychology journals representing 13 subfields (Fig. S1 in the Supplemental Material). We asked respondents about the frequency of statistical review in their journal, the nature of their statistical reviewers and how they are chosen, the procedures and outcomes of statistical review, their ability and willingness to use statistical review, and their perception of the value of statistical review.

An unexpected observation both complicated interpretation of the data and motivated this commentary; 17 (44%) respondents stated that no additional specialized statistical review was warranted and that regular peer reviewers are both capable of evaluating and expected to evaluate the statistical aspects of submitted manuscripts (see Results in the Supplemental Material). These views contrast starkly with those of biomedical editors and statisticians (Wasserstein & Lazar, 2016), who almost universally accept the notion that statistical errors or suboptimal analyses can go undetected by regular peer review, and that specialized and targeted statistical review is required (Hardwicke & Goodman, 2019).

Does Psychology Need Statistical Review?

Researchers have highlighted a litany of statistical ailments that afflict the psychology literature, ranging from simple reporting errors to wholesale misunderstanding and misapplication of fundamental statistical concepts and techniques (see Table 1). One striking example is the pervasive problem of inadequate statistical power that persists in several domains of psychology. Many published psychology studies have such small sample sizes that statistical tests are unlikely to be sufficiently powered to detect plausible effects (e.g., Button et al., 2013; Cohen, 1962; Fraley & Vazire, 2014; Sedlmeier & Gigerenzer, 1989; Stanley, Carter, & Doucouliagos, 2018; Szucs & Ioannidis, 2017; Vankov, Bowers, & Munafò,

Table 1. Statistical Ailments in the Published Psychology (and Related) Literature, With References Providing Further Detail and Empirical Evidence

Issue	References
Selection of analyses	
Using inappropriate statistical models (e.g., using metric models to analyze ordinal data, using an independent-samples *t* test in a repeated measures design, neglecting model assumptions)	Ernst and Albers (2017); Liddell and Kruschke (2018)
Insufficient sample size to achieve reasonable statistical power (see the main text for details)	Button et al. (2013); Cohen (1962); Fraley and Vazire (2014); Maxwell (2004); Sedlmeier and Gigerenzer (1989); Smaldino and McElreath (2016); Stanley, Carter, and Doucouliagos (2018); Szucs and Ioannidis (2017); Vankov, Bowers, and Munafò (2014)
Circular analysis (e.g., attempting to correlate brain-activity measures with personality measures after selecting from the former only data that have surpassed a threshold—also known as "double dipping")	Fiedler (2011); Vul, Harris, and Winkielman (2009)
Implementation of analyses	
Failure to account for multiplicity (e.g., multiple comparisons, optional stopping)	Cramer et al. (2016); John, Loewenstein, and Prelec (2012); Simmons, Nelson, and Simonsohn (2011)
Unjustified exclusion of outliers (e.g., excluding data points ad hoc in a way that makes outcomes more favorable to the hypothesis under scrutiny)	Bakker and Wicherts (2014); John et al. (2012)
Incorrect calculation of effect sizes (e.g., using erroneous formulas)	Hardwicke et al. (2018)
Falsification of data	Fanelli (2009); John et al. (2012); Simonsohn (2013)
Reporting of analyses	
Inconsistencies (e.g., the reported degrees of freedom, test statistic, and *p* value are incompatible; the reported means for integer data, sample size, and number of items are incompatible)	Bakker and Wicherts (2011); Brown and Heathers (2017); Nuijten, Hartgerink, van Assen, Epskamp, and Wicherts (2016)
Misleading or suboptimal graphical presentation (e.g., inappropriate truncation of the *y*-axis, misidentification or nonidentification of error bars, no display of distributional information)	Lane and Sándor (2009)
Incomplete or unclear specification of the design or analysis (e.g., not identifying statistical procedures, ambiguously describing experimental units, not reporting data exclusions)	Hardwicke et al. (2018); Lazic, Clarke-Williams, and Munafò (2018)
Incomplete reporting of outcomes (e.g., not reporting effect sizes, interval estimates, or standard deviations)	Counsell and Harlow (2017); Cumming et al. (2007); Tressoldi, Giofré, Sella, and Cumming (2013)
Selective reporting (e.g., reporting only experiments or outcomes that achieved statistical significance)	Franco, Malhotra, and Simonovits (2015); John et al. (2012); Simmons et al. (2011)
Interpretation of analyses	
Overstating conclusions when there is low evidential value	Wetzels et al. (2011)
Presenting post hoc hypotheses as if they were specified a priori (sometimes called HARKing)	John et al. (2012); Kerr (1998); Wagenmakers, Wetzels, Borsboom, Maas, and van der Kievit (2012)
Incorrectly concluding that a nonsignificant outcome means that there is "no effect"	Dienes (2014); Finch, Cumming, and Thomason (2001); Sedlmeier and Gigerenzer (1989)
Assuming that the difference between significant and not significant is itself significant or analyzing interactions erroneously	Nieuwenhuis, Forstmann, and Wagenmakers (2011); Gelman and Stern (2006)

2014). Smaldino and McElreath (2016) examined 44 studies of statistical power in the social and behavioral sciences and found that the average power to detect small-size effects (d = 0.2) was very low (M = 0.24, assuming α = .05). Moreover, there has generally been no increase in power over time despite repeated calls to address the issue (Button et al., 2013; Cohen, 1962; Sedlmeier & Gigerenzer, 1989; but see Sassenberg & Ditrich, 2019). Because statistical power is a function of multiple factors, the problem may be less severe in domains (such as psychophysics) that commonly feature low intrasubject variability, within-subjects designs, and multiple measurement trials per subject (Rouder & Haaf, 2017). Inadequate statistical power, coupled with publication bias, can lead to inflated effect-size estimates and increases the likelihood of false negatives and false discoveries (Button et al., 2013; Fraley & Vazire, 2014; Ioannidis, 2005). Survey evidence and examination of articles' Method sections suggests that many psychologists choose sample sizes on the basis of typical practice in their domains of research rather than formal power analysis (Sedlmeier & Gigerenzer, 1989; Vankov et al., 2014). As these domain experts are also training the next generation of scientists and scrutinizing their colleagues' work during the peer-review process, a self-reinforcing cycle of suboptimal practice may follow. Independent statistical review that focuses on issues like those listed in Table 1 (as occurs at biomedical journals; e.g., Cobo et al., 2007; Gore, Jones, & Thompson, 1992) could help to break such cycles.

The pervasiveness of statistical ailments in the published literature suggests that peer review in psychology journals is not sufficient to identify and minimize those problems. Quantitative training programs in psychology are typically slow to incorporate contemporary developments, avoid advanced topics, and provide only superficial treatment of fundamental statistical concepts (Aiken, West, & Millsap, 2008; Aiken, West, Sechrest, & Reno, 1990). Much quantitative training in psychological science neglects historical and philosophical foundations (Gigerenzer, 2004, 2018), proliferating confusion about core statistical concepts and facilitating widespread adoption of suboptimal practices (Wasserstein & Lazar, 2016). Statistical misconceptions are prevalent among instructors and deeply embedded in mainstream research-methods curricula (Brewer, 1985; Haller & Krauss, 2002; for a review, see Gigerenzer, 2018). Some research practices taught to undergraduates are now recognized as questionable (Bem, 2004; Wagenmakers, Wetzels, Borsboom, Maas, & van der Kievit, 2012).

Why Is Statistical Review Used in Medicine?

Leading biomedical journals have been adopting statistical review and refining their policies since the 1970s (Altman, 1982, 1994, 1998; Smith, 2005). Most biomedical-journal editors in our survey (Hardwicke & Goodman, 2019) indicated that they believed statistical review provides substantial incremental value beyond regular peer review and results in important changes to manuscripts around 60% of the time—even though many biomedical articles have Ph.D.-level methodologists among the authors. This view is supported by empirical work evaluating leading medical journals, including *The BMJ, The Lancet*, and *Annals of Internal Medicine*, which has consistently indicated that statistical review can play an important role in improving manuscript quality (Gardner & Bond, 1990; Goodman, Berlin, Fletcher, & Fletcher, 1994; Gore et al., 1992; Prescott & Civil, 2013; Schor & Karten, 1966). In a 2017 *Annals of Internal Medicine* survey of 337 corresponding authors of research published between 2012 and 2016, 57% reported a moderate or large increase in their article's overall quality as a result of the statistical editorial process; only 15% reported "no" impact, and only 2% reported a "negative" impact. In addition, 58% reported making considerable effort to respond to statistical comments, and 54% felt that such effort was "definitely" worthwhile (Stack et al. 2017).

To our knowledge, there has been only one randomized control trial designed to evaluate the effectiveness of statistical review (Cobo et al., 2007). That study, conducted at the biomedical journal *Medicina Clinica*, involved 115 articles, 16 of which were ultimately not published. The addition of statistical review to regular peer review led to small quality increases for all but 3 of 36 assessment criteria, resulting in overall modest but discernible improvements in manuscript quality. Although this improvement could have been due to simply adding a reviewer, it was the statistical aspects of the manuscripts that improved most.

Providing statistical guidelines for authors makes a journal's expectations transparent and may help to improve statistical practice (Bailar & Mosteller, 1988; Smith, 2005). Many psychology journals indicate that authors should adhere to statistical-reporting guidelines, such as those from the American Psychological Association (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008; Wilkinson & APA Task Force on Statistical Inference, 1999) and the Psychonomic Society (2019).

The evidence for the effectiveness of statistical guidelines in biomedical journals is mixed (Dexter & Shafer, 2017). In psychology, the introduction of journal-specific statistical guidelines at the journal *Psychological Science* was associated with a number of modest improvements in statistical reporting (Giofrè, Cumming, Fresc, Boedker, & Tressoldi, 2017). Such guidelines are most effective if enforced by a reviewing editor (Dexter & Shafer, 2017) and may improve the efficiency, completeness, and standardization of statistical review, but they are unlikely to supplant expert statistical review (Altman, 1998).

In summary, there is a reasonable body of evidence to suggest that specialized statistical review in biomedicine has been effective in preventing many serious analytic and inferential errors from reaching the published literature. Could psychology journals improve the validity and reproducibility of their content by adopting a similar model?

How Would Statistical Review Work in Psychology?

In the biomedical domain, there is no single model for statistical review (Hardwicke & Goodman, 2019), and policies have evolved gradually (Altman, 1998; Cobo et al., 2007; Gardner & Bond, 1990; Goodman et al., 1994; Gore et al., 1992; Prescott & Civil, 2013; Schor & Karten, 1966; Smith, 2005). Drawing from that experience, we address four key logistical issues: Who should conduct statistical review, which manuscripts should undergo statistical review, at what stage should statistical review be performed, and how should statistical review be incorporated into the editorial process?

Who should conduct statistical review?

Psychology statistical reviewers do not necessarily need to be statisticians per se, but should have advanced (Ph.D.-level) quantitative training (Goodman et al., 1998; Hardwicke & Goodman, 2019). Ideally, they should understand the terminology, conventions, and practices of psychology research. The majority of psychology-journal editors responding to our survey reported that difficulty in finding appropriate reviewers affected their willingness to conduct statistical review (see Fig. S3 in the Supplemental Material). It is not clear whether this difficulty reflected a lack of potential reviewers or problems identifying them.

The number of statistical reviewers required for a journal will depend on the model of statistical review employed. Just over half of the biomedical journals we surveyed indicated that they typically relied on around two statistical experts on their internal editorial teams to conduct all of the statistical review (Hardwicke & Goodman, 2019), although the largest journals tended to have more. Just over a third relied on a pool of external reviewers, with a median size of 11 members. In our psychology survey, the majority of respondents who reported using statistical review indicated that their statistical reviewers were typically identified on an ad hoc basis (58%; see Fig. S4a in the Supplemental Material). Many relied on from 1 to 40 editorial-team members (*Mdn* = 20, plus 3 missing responses), although it was unclear whether these individuals had specialized statistical expertise. Only one respondent indicated that there was a predesignated pool of external statistical reviewers, consisting of 25 members.

A starting point for psychology journals could be to recruit one statistical expert to serve on the editorial board or be retained as a regular consultant. If the expert is not someone who would see this as a professional service or a vehicle for career advancement, compensation might be required. Whereas about half of biomedical journals pay their statistical reviewers, only one respondent in our psychology survey did so (Fig. S4c in the Supplemental Material).

Which manuscripts should undergo statistical review?

Optimally, all likely-to-be-accepted manuscripts with relevant statistical content should undergo statistical review (George, 1985; Schor & Karten, 1966). In our psychology survey (see Fig. S2 in the Supplemental Material), 15 (38%) respondents indicated that statistical review was used for all relevant articles, and 15 (38%) respondents indicated that statistical review was rare (≤ 10% of articles). However, free-text comments indicated that at least 8 of the 15 who said all manuscripts received such review did not differentiate it from regular peer review; some of these editors indicated that peer reviewers had sufficient statistical training.

If all articles cannot be statistically reviewed, editors have to prioritize. Smith (2005) noted that it took 5 to 10 years for *The BMJ* to reach the point where all published articles with a statistical component were undergoing statistical review. The *Annals of Internal Medicine* had one statistical reviewer in the early 1980s and added a second in 1987. The team grew steadily over the ensuing 30 years to its current size of 10. The journal *Psychological Science* has recruited a pool of 6 statistical advisors who can be called upon by the journal's editors at their discretion (Association for Psychological Science, 2016).

Targeting manuscripts with complex methods for statistical review makes some sense, but a number of commentators in the biomedical domain have noted that routine statistical analyses tend to be the most problematic (Schor & Karten, 1966; Smith, 2005). Sophisticated analyses may be conducted by individuals with more statistical expertise (Schor & Karten, 1966). Many of the statistical ailments in the psychology literature relate to foundational issues, not advanced techniques. Consequently, the most impactful contribution of statistical review might come from evaluating what appear to be routine analyses.

At what stage should statistical review be performed?

An important question for journals is, at what stage of the publication process should manuscripts undergo statistical review? In our survey of biomedical journals, the majority of respondents indicated that statistical review was either solicited at the same time as regular peer review (36%) or after regular peer review and before a provisional acceptance decision (27%). In our psychology survey, although the majority of respondents (71%) indicated that statistical review was solicited at the same time as regular peer review (Fig. S5c), many did not differentiate between regular and statistical review. The model will ultimately be journal-specific, dependent on the journal's capacity for statistical review.

How should statistical review be incorporated into the editorial process?

How editors should incorporate the input of statistical reviewers is an important issue, particularly for journals unused to such review. Smith (2005) described the slow process of mutual education that had to occur at *The BMJ*:

> We worried that the gulf between medical editors and statisticians with no knowledge of medical research would be unbridgeable. . . . In the early days we made the mistake of thinking that statistics was a much more exact science than clinical research and that we had to go along with exactly what the statisticians advised. Eventually we learnt that there was room for negotiation over what was acceptable . . . recognizing the inevitable trade-offs between statistical purity [and] what can actually be done in clinical research. . . . (p. 2)

Smith's observations illustrate that effective statistical review requires not only the addition of a statistical reviewer, but also "cross-cultural" education and communication, which takes time. The reviewers need to understand and absorb the values of the research community they are serving, and that community, and the editors, needs to understand how the changes requested by such reviewers are improving the validity of its research. External statistical reviewers who are not part of the journal can make unrealistic requests, which must be adjudicated or modified by an internal editor. Statistical experts directly incorporated into the editorial process absorb journal and disciplinary norms and are also able to educate editors.

Statistical Review, Open Science, and Metaresearch

Psychological science is in the midst of a credibility revolution (Nelson et al., 2018; Vazire, 2018), and this is an opportune time for journal editors to consider adoption of statistical review. There is growing awareness that the credibility of scientific claims depends on transparent reporting (Klein et al., 2018; Munafò et al., 2017). Statistical review is likely to be most effective when reviewers have access to all of the raw research artifacts (materials, data, analysis scripts, and preregistered protocols when relevant), which enable a fully informed assessment. Having access to data, and ideally analysis scripts, enables verification of analytic reproducibility (Hardwicke et al., 2018; Sakaluk, Williams, & Biernat, 2014) and assessment of analytic robustness (LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018; Localio et al., 2018 Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016), and it can facilitate detection of fraud (Simonsohn, 2013; Smith, 2005). Research materials can convey statistically relevant information about data collection, and availability of survey instruments can help reviewers raise questions about psychometric issues (McPherson & Mohr, 2005). Preregistration of study protocols (Nosek, Ebersole, DeHaven, & Mellor, 2018) could facilitate identification of questionable research practices such as *p*-hacking and HARKing (i.e., Hypothesizing After the Results are Known; O'Boyle, Banks, & Mulé, 2017). At some psychology journals, practices such as data sharing are already being actively encouraged (e.g., *Psychological Science*; Kidwell et al., 2016) or mandated (e.g., *Cognition*; Hardwicke et al., 2018; also see Nuijten et al., 2017).

Statistical review might be enhanced by the use of computer algorithms to automatically screen for and flag potential errors in submitted manuscripts. The free software *statcheck* (http://statcheck.io/), for example, can automatically extract certain statistical outcomes reported in American Psychological Association style and check the internal consistency of *p* values, test statistics, and degrees of freedom (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Semiautomated tools like this require little resource investment and could reduce the burden of statistical review. One psychology editor who responded to our survey already required authors to submit statcheck reports with their manuscripts.

Increasing input from quantitative experts before a study begins could be an especially impactful approach to improving the quality of statistical scaffolding. A growing number of psychology journals, such as *Nature Human Behaviour* and *Collabra: Psychology*, are adopting the Registered Report format, which involves peer review of study protocols before the study has begun (Chambers, 2013; Hardwicke & Ioannidis, 2018). The contribution of expert statistical reviewers in the early stages of a research project could well be the most effective and efficient use of their time.

Finally, psychologists not only are driving the development of new reform initiatives, but also are conducting empirical investigations to evaluate the effectiveness of these initiatives in order to iteratively improve upon them (e.g., Hardwicke & Ioannidis, 2018; Hardwicke et al., 2018; Kidwell et al., 2016; Nuijten et al., 2017). These exercises in metaresearch (Hardwicke et al., 2019; Ioannidis, Fanelli, Dunne, & Goodman, 2015) should be extended to statistical review. A series of prospectively registered randomized control trials designed to evaluate various models of statistical review would be a valuable tool for gathering evidence relevant to this issue.

Conclusion

In this article, we have advocated that psychology journals consider adopting specialized statistical review to complement regular peer review. We have been partly informed by the results of a survey of psychology-journal editors; however, given the small number of respondents, likelihood of self-selection bias, and reliance on self-report, only tentative inferences can be drawn from these data. Our arguments are mainly based on the apparent benefits of statistical review in the biomedical domain and the documented statistical problems pervading the psychology research literature. We contend that there is sufficient evidence to support pilot testing expert statistical review in psychology journals, with concomitant monitoring and evaluation.

Statistical review will not cure all of psychology's statistical ailments, just as it is no panacea in biomedicine. The most effective antidote is likely to involve efforts to improve statistical competence among psychology researchers (Aiken et al., 2008), and to promote more open science, which would enable more effective postpublication review. This will require nontrivial reforms in training curricula and normative structures surrounding design, analysis, and inference. If psychology is to break free of problematic statistical rituals (Salsburg, 1985) and make better use of the analysis toolbox (Gigerenzer, 2014, 2018), it will require an infusion of fresh thinking from well-trained quantitative experts at all stages of the teaching, research, funding, and publication pipeline.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

T. E. Hardwicke and S. N. Goodman designed the survey. T. E. Hardwicke conducted the survey and analyzed the survey data. T. E. Hardwicke, M. C. Frank, S. Vazire, and S. N. Goodman wrote the manuscript.

ORCID iDs

Tom E. Hardwicke https://orcid.org/0000-0001-9485-4952
Michael C. Frank https://orcid.org/0000-0002-7551-4378

Acknowledgments

We thank Lisa Ann Yu for assistance collecting journals' contact details and Daniele Fanelli for discussions about the survey design. We are grateful to all the respondents for taking the time to complete the survey.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was enabled in part by a general support grant to the Meta-Research Innovation Center at Stanford (METRICS) from the John and Laura Arnold Foundation. The Meta-Research Innovation Center Berlin (METRIC-B) is supported by a grant from the Einstein Foundation and Stiftung Charité.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245919858428

Open Practices

Open Data: https://osf.io/nquws/files/
Open Materials: https://osf.io/tmah8/files/, https://osf.io/4zurk/files/
Preregistration: no
All data and materials, including analysis scripts, have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/nquws/files/, https://osf.io/tmah8/files/, and https://osf.io/4zurk/files/, respectively. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245919858428. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges

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. *American Psychologist*, *63*, 32–50.

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist*, *45*, 721–734.

Altman, D. G. (1982). Statistics in medical journals. *Statistics in Medicine*, *1*, 59–71.

Altman, D. G. (1994). The scandal of poor medical research. *BMJ*, *308*, 283–284.

Altman, D. G. (1998). Statistical reviewing for medical journals. *Statistics in Medicine*, *17*, 2661–2674.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839–851.

Association for Psychological Science. (2016, March). Meet *Psychological Science*'s new statistical advisors. *Observer*. Retrieved from https://www.psychologicalscience.org/observer/meet-psychological-sciences-new-statistical-advisors

Aust, F., & Barth, M. (2019). papaja: Prepare APA journal articles with R Markdown (R package Version 0.1.0.98) [Computer software]. Retrieved from https://github.com/crsh/papaja

Bailar, J. C., & Mosteller, F. (1988). Guidelines for statistical reporting for medical journals: Amplifications and explanations. *Annals of Internal Medicine*, *108*, 266–273.

Baker, M. (2016). Is there a reproducibility crisis? A *Nature* survey lifts the lid on how researchers view the crisis rocking science and what they think will help. *Nature*, *533*, 452–454.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–678.

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples *t* tests: The power of alternatives and recommendations. *Psychological Methods*, *19*, 409–427.

Bem, D. J. (2004). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger, III (Eds.), *The compleat academic: A career guide* (2nd ed., pp. 185–219). Washington, DC: American Psychological Association.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational and Behavioral Statistics*, *10*, 252–268.

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM test. *Social Psychological & Personality Science*, *8*, 363–369.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 1–12.

Chambers, C. D. (2013). *Registered Reports*: A new publishing initiative at *Cortex*. *Cortex*, *49*, 609–610.

Cobo, E., Selva-O'Callagham, A., Ribera, J.-M., Cardellach, F., Dominguez, R., & Vilardell, M. (2007). Statistical reviewers improve reporting in biomedical articles: A randomized trial. *PLOS ONE*, *2*(3), Article e332. doi:10.1371/journal.pone.0000332

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Counsell, A., & Harlow, L. L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology/Psychologie Canadienne*, *58*, 140–147.

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., . . . Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*, 640–647.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., . . . Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, *18*, 230–232.

Dexter, F., & Shafer, S. L. (2017). Narrative review of statistical reporting checklists, mandatory statistical editing, and rectifying common problems in the reporting of scientific articles. *Anesthesia & Analgesia*, *124*, 943–947.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, Article 781. doi:10.3389/fpsyg.2014.00781

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice—A systematic review of common misconceptions. *PeerJ*, *5*, Article e3323. doi:10.7717/peerj.3323

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE*, *4*(5), Article e5738. doi:10.1371/journal.pone.0005738

Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163–171.

Finch, S., Cumming, G., & Thomason, N. (2001). Colloquium on effect sizes: The roles of editors, textbook authors, and the publication manual: Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement*, *61*, 181–210.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS ONE*, *9*(10), Article e109019. doi:10.1371/journal.pone.0109019

Franco, A., Malhotra, N., & Simonovits, G. (2015). Underreporting in psychology experiments. *Social Psychological & Personality Science*, *7*, 8–12.

García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology*, *3*, Article 325. doi:10.3389/fpsyg.2012.00325

Gardner, M. J., & Bond, J. (1990). An exploratory study of statistical assessment of papers published in the British Medical Journal. *JAMA*, *263*, 1355–1357.

Gelman, A., & Stern, H. (2006). The difference between significant and not significant is not itself statistically significant. *The American Statistician*, *60*, 328–331.

George, S. L. (1985). Statistics in medical journals: A survey of current policies and proposals for editors. *Pediatric Blood & Cancer*, *13*, 109–112.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*, 198–218.

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PLOS ONE*, *12*(4), Article e0175583. doi:10.1371/journal.pone.0175583

Goodman, S. N., Altman, D. G., & George, S. L. (1998). Statistical reviewing policies of medical journals. *Journal of General Internal Medicine*, *13*, 753–756.

Goodman, S. N., Berlin, J., Fletcher, S. W., & Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at *Annals of Internal Medicine*. *Annals of Internal Medicine*, *121*, 11–21.

Gore, S. M., Jones, G., & Thompson, S. G. (1992). The Lancet's statistical review process: Areas for improvement by authors. *The Lancet*, *340*, 100–102.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*(1), 1–20.

Hardwicke, T. E., & Goodman, S. N. (2019). *A survey of statistical reviewing policies at leading biomedical journals*. Manuscript in preparation.

Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour*, *2*, 793–796.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., . . . Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, *5*(8), Article e180448. doi:10.1098/rsos.180448

Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. P. A. (2019). Calibrating the scientific ecosystem through meta-research. *MetaArXiv*. doi:10.31222/osf.io/krb58

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), Article e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLOS Biology*, *13*(10), Article e1002264. doi:10.1371/journal.pbio.1002264

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

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*, 196–217.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., . . . Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*(5), Article e1002456. doi:10.1371/journal.pbio.1002456

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., . . . Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*, 2474–7394.

Lane, D. M., & Sándor, A. (2009). Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychological Methods*, *14*, 239–257.

Lazic, S. E., Clarke-Williams, C. J., & Munafò, M. R. (2018). What exactly is '*N*' in cell culture and animal experiments? *PLOS Biology*, *16*(4), Article e2005282. doi:10.1371/journal.pbio.2005282

LeBel, E. P., McCarthy, R., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*, 389–402.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Localio, A. R., Goodman, S. N., Meibohm, A., Cornell, J. E., Stack, C. B., Ross, E. A., & Mulrow, C. D. (2018). Statistical code to support the scientific story. *Annals of Internal Medicine*, *168*, 828–829.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

McPherson, J., & Mohr, P. (2005). The role of item extremity in the emergence of keying-related factors: An exploration with the Life Orientation Test. *Psychological Methods*, *10*, 120–131.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 21. doi:10.1038/s41562-016-0021

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 1511–1534.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Human Behaviour*, *14*, 1105–1109.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA*, *115*, 2600–2606.

Nuijten, M. B., Borghuis, J., Veldkamp, C. L. S., Dominguez-Alvarez, L., Van Assen, M. A. L. M., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology*, *3*, Article 31. doi:10.1525/collabra.102

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226.

O'Boyle, E. H., Jr., Banks, G. C., & Gonzalez-Mulé, E. (2017). The Chrysalis Effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, *43*, 376–399.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Prescott, R. J., & Civil, I. (2013). Lies, damn lies and statistics: Errors and omission in papers submitted to INJURY 2010–2012. *Injury*, *44*, 6–11.

Psychonomic Society. (2019). *Statistical guidelines*. Retrieved from https://www.psychonomic.org/page/statisticalguidelines

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

Rouder, J. N., & Haaf, J. M. (2017). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*, 19–26.

Sakaluk, J., Williams, A., & Biernat, M. (2014). Analytic review as a solution to the misreporting of statistical results in psychological science. *Perspectives on Psychological Science*, *9*, 652–660.

Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician*, *39*, 220–223.

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*, 107–114.

Schor, S., & Karten, I. (1966). Statistical evaluation of medical journal manuscripts. *JAMA*, *195*, 1123–1128.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

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

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875–1888.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), Article 160384. doi:10.1098/rsos.160384

Smith, R. (2005). Statistical review for medical journals, journal's perspective. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (2nd ed.). Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a17141

Sox, H. C. (2009). Medical journal editing: Who shall pay? *Annals of Internal Medicine*, *151*, 68–69.

Stack, C., Ludwig, A., Localio, A. R., Meibohm, A., Guallar, E., Wong, J., . . . Laine, C. (2017, September). *Authors' assessment of the impact and value of statistical review in a general medical journal: 5-year survey results*. Poster presented at the Eighth International Congress on Peer Review and Scientific Publication. Retrieved from https://peerreviewcongress.org/prc17-0202

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*, 1325–1346.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), Article e2000797. doi:10.1371/journal.pbio.2000797

Tressoldi, P. E., Giofré, D., Sella, F., & Cumming, G. (2013). High impact = high statistical standards? Not necessarily so. *PLOS ONE*, *8*(2), Article e56180. doi:10.1371/journal.pone.0056180

Vankov, I., Bowers, J., & Munafò, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology*, *67*, 1037–1040.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*, 411–417.

Vul, E., Harris, C. R., & Winkielman, P. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*, 291–298.

Wilkinson, L., & APA Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Xie, Y. (2018). knitr: A general-purpose package for dynamic report generation in R (R package Version 1.23) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/knitr/