Published in final edited form as:

Basic Appl Soc Psych. 2017 ; 39(5): 239–246. doi:10.1080/01973533.2017.1336093.

Results Blind Science Publishing

Joseph J. Locascio

Harvard Medical School

Abstract

Problems in science publishing involving publication bias, null hypothesis significance testing (NHST), and irreproducibility of reported results have been widely cited. Numerous attempts to ameliorate these problems have included statistical methods to assess and correct for publication bias, and recommendation or development of statistical methodologies to replace NHST where some journals have even instituted a policy of banning manuscripts reporting use of NHST. In an effort to mitigate these problems, a policy of “results blind evaluation” of manuscripts submitted to journals is recommended, in which results reported in manuscripts are given no weight in the decision as to the suitability of the manuscript for publication. Weight would be given *exclusively* to (a) the judged *importance* of the research question addressed in the study, typically conveyed in the Introduction section of the manuscript, and (b) the quality of the *methodology* of the study, including appropriateness of data analysis methods, as reported in the Methods section. As a practical method of implementing such a policy, a two-stage process is suggested whereby the editor initially distributes only the Introduction and Methods sections of a submitted manuscript to reviewers for evaluation and a provisional decision regarding acceptance or rejection for publication is made. A second stage of review follows in which the complete manuscript is distributed for review but only if the decision of the first stage is for acceptance with no more than minor revision.

Problems in science publishing

Controversy regarding certain problems in science publishing, especially in the behavioral and medical sciences, has recently become intensified. These problems are (a) publication bias, that is, a widely acknowledged tendency among scientific journals toward preferentially publishing positive research findings; (b) problems associated with the use of null hypothesis significance testing (NHST); and (c) reported irreproducibility of published research findings.

Publication bias

In this article *publication bias* means the widely recognized practice in which, all else being equal, reports of null findings, or statistically nonsignificant results as determined by NHST, are not as likely as those of positive findings to be published in scientific journals, either

CONTACT Joseph J. Locascio, jlocascio@partners.org, Harvard Medical School, Neurology, Boston, MA, USA.

ORCID

Joseph J. Locascio <http://orcid.org/0000-0003-3439-1209>

because manuscripts reporting such are not as likely to be accepted for publication by journals (Fanelli, 2010; Van Assen, Van Aert, Nuijten, & Wicherts, 2014) or because of an inhibition among investigators in even submitting such reports to journals because of this assumed bias at the journals (the “file drawer” problem discussed by Rosenthal, 1979). (Publication can be biased in other ways, but arguably not in as common or pernicious a manner as this; Coburn & Vevea, 2015). A study with nonsignificant findings is often regarded by its investigators as a “failure”, instead of potentially providing valid and useful information about the absence or negligibility of a relation, assuming the study was methodologically sound with adequate sample size. Various statistical and graphical methods have been devised to assess whether publication bias exists especially in the context of meta-analyses (e.g., the “funnel plot” and similar methods; Wang & Bushman, 1998), and evidence for it is commonly found (Kühberger, Fritz, & Scherndl, 2014; Van Assen, Van Aert, & Wicherts, 2015).

A potential negative consequence of publication bias can be easily illustrated with a simplified example. If, for example, a new medication, which actually has no effect whatever on a disease condition that it is intended to treat, is tested for effectiveness in 100 independent studies across the world, each using let’s say a one-tail NHST with $\alpha = 0.05$, approximately five of these studies will likely obtain a significant finding, purely as a result of chance. That’s what a Type I error of .05 means; in the long run, if the null hypothesis is true, five of 100 independent tests of it will show a significant effect. Then, if only reports from these five studies alone are submitted and/or accepted for publication because they were the only ones showing significant results, the research community will be misled into thinking that the effectiveness of the drug has widespread and unanimously confirmed validation. (Results from two-tailed tests are almost as bad assuming counterintuitive results are self-censored as erroneous.) This example is admittedly extreme, but it’s easy to see that to the degree that a true effect is negligible and to the degree that publication bias is operative at the author submission and/or journal acceptance stage, the literature will be contaminated with a misleadingly higher proportion of false positive findings than would occur by chance alone if there were no such biases.

An even more pernicious problem associated with publication bias is the temptation that some researchers may feel to doctor their results to show a positive finding in order to get the publications they believe are vital to their career advancement. In the context of NHST, this might mean “fudging the data” to get the p below .05. Moreover, some have argued that there are many ways that less deliberate, more unconscious processes and choices may be at play in how researchers collect and analyze their data in order to obtain the significant findings they feel they need to secure publication (e.g., “ p hacking” and “opportunistic biases”; DeCoster, Sparks, Sparks, Sparks, and Sparks, 2015). A vicious cycle can develop of somewhat innocent publication bias inducing a less innocent kind of bias.

There have been some statistical methods proposed and used to try to correct for publication bias, for example, as regards meta-analysis (Citkowitz & Vevea, 2017), but there seems to be a widespread assumption that nothing can be done about the bias, or worse, that it is a good practice and nothing *should* be done about it (e.g., see Schwarzkopf, 2015), that is,

given limited space in journals, we need to know only about the studies that found something, not the ones that “didn’t find anything.”

Null hypothesis significance testing

Controversy concerning the common inferential statistical practice of NHST, especially in psychological research, has recently peaked. Briefly, the NHST assumes a null hypothesis (H_0) of no effect in a referent population and determines whether representative sample data (D) deviates from H_0 to the point of having such a small probability of occurrence given H_0 (the “ p value” or $p(D|H_0)$, conventionally less than .05), that H_0 is rejected and some prestatated alternate hypothesis (H_A) more consistent with the data is tentatively accepted as true. The use of NHST has been ubiquitous for decades especially in the behavioral, social, and medical sciences. Obtaining a “significant” NHST result has become virtually synonymous with a “positive” finding for a study and a “nonsignificant” result with a “negative” or “null” finding. Further, because of the publication bias among journals to give preference to reports of positive findings, just noted, the use of NHST has effectively become a widespread tacitly assumed requirement of much of research with its favorable result often assumed to be a sine qua non for publication. Effectively, NHST has been singled out and consensually given the official role of the operative gatekeeper of much of publishable truth in psychology, and much of science in general.

Despite this central role of NHST in research, serious and valid concerns about its use and misuse have been cited by statisticians for decades, especially within the field of psychology (see, e.g., Cohen, 1994; Hunter, 1997; Seife, 2015, for a sampling of criticisms of NHST). For example, the NHST is said to ask the wrong question (i.e., about what $p(D|H_0)$ is instead of the more relevant $p(H_0|D)$ or $p(H_A|D)$), more often than not, to give the wrong answer (frequently false negatives and positives, or even if true positives, with a high biased effect size leading to difficulty of replication; Schmidt, 1996), to be misunderstood as providing $p(H_0|D)$ or a direct index of effect size or substantive importance, to be ambiguous regarding the domain over which to control for false positives (i.e., Type I error; Trafimow & Earp, 2017), among many other criticisms. More highly regarded alternatives to the NHST have been widely recommended (e.g., confidence intervals, effect size estimates, Bayesian statistics, meta-analyses, graphical analysis, etc.), though none of them has come close to replacing the NHST as the workhorse of statistical inference. These issues gained especially wide attention in early 2015, when editors of the journal *Basic and Applied Social Psychology* (*BASP*) decided to address NHST problems by instituting an outright ban on reporting of NHST results and related statistics (e.g., confidence intervals, especially when used as an indirect NHST) in that journal (Trafimow & Marks, 2015). The American Statistical Association recently even felt it necessary to publish a policy statement critical of widespread misuse of the NHST and p values (Wasserstein & Lazar, 2016).

Irreproducibility of reported findings

There has been so much recent concern about a failure to replicate findings in scientific reports, especially but not exclusively in psychological science, that the problem has been widely referred to as a “replication crisis” (Ioannidis, 2005; Johnson, Payne, Wang, Asher, & Soutrik, 2017; Maxwell, Lau, & Howard, 2015; Winerman, 2016). A recent large-scale

attempt to replicate previously reported statistically significant findings in psychology found only about one third of them to replicate (Open Science Collaboration, 2015). The reasons for these failures have been heavily debated with publication bias being commonly alleged to be partly responsible, which includes biases at the investigator level whereby various conscious and unconscious stratagems are employed to get significant results (e.g., “p-hacking,” “opportunistic biases”; DeCoster et al., 2015). It has also been claimed that the bias of publishing only results with NHST p values below the .05 cutoff specifically has made much failure to replicate inevitable given statistical regression (i.e., “regression to the mean”; Trafimow & Earp, 2017).

Results blind manuscript evaluation

As a method of mitigating the aforementioned problems, I suggest that academic psychology, social science, and medical journals consider instituting a policy of results blind manuscript evaluation (RBME) of the suitability of submitted manuscripts for publication (Locascio, 1999, 2011, 2015). (This would pertain only to reports of empirical research for single studies or meta-analysis, not papers on purely methodological/data analysis issues, psychometric/biometric reports of development of measurement instruments, literature review papers, editorial-opinion papers, letters to the editor, and such.) Although the author came to this conclusion independently (Locascio, 1999), the suggestions proposed here are not claimed to be entirely new and original. A few previously proposed solutions in various fields of science have at one time or another recommended essentially the same thing or something close to it, via published articles, editorials, and letters to editors (Armstrong, 1997; Colom & Vieta, 2011; DeCoster et al., 2015; Glymour & Kawachi, 2005; Greve, Broder, & Erdfelder, 2013; Kupfersmid, 1988; Newcombe, 1987; Smulders, 2013). Some have on occasion informally suggested something similar in online blogs (Hanson, 2007, 2010), and it’s possible others have done so also. Further, there are a few journals that currently employ editorial policies or submission options similar to RBME. For example, some journals now state in their guidelines for authors that they will accept “null findings” (e.g., *BASP*; Trafimow, 2014). This would be implicit in RBME. However, there appears to be no widespread recognition of RBME as a possible solution to the general controversies concerning publication bias, NHST, irreproducibility of results, and related issues, and there is certainly no widespread use of these proposed methods.

To be more specific, by RBME I mean that in deciding whether to publish a manuscript, journal reviewers and editors would give no weight to reported results in making this judgment. Reports would be judged *only* on the basis of perceived (a) *importance* of the substantive issues addressed, which is generally communicated in the Introduction section of the manuscript, and (b) soundness of the *methodology* employed, as conveyed in the Methods section, which would include, among other things, appropriateness of: materials; measurement methods with, for example, upfront reliability computations; subject selection; study layout/experimental design; proposed data analysis techniques; and sample size. (Note that a good assessment of a study’s reported methodology would almost certainly require some substantive knowledge of the subject area, not merely expertise with abstract issues of study design, data analysis techniques, etc.) Regarding appropriate data analysis methodology, many proposed methods recommended as alternatives to NHST could

potentially be employed, instead of or in addition to NHST. In my view, use of NHST, or reporting of a p value as a continuous estimate of $p(D|H_0)$, if applicable, and correctly used and interpreted, could conceivably be permissible as providing one possible, optional, supplementary piece of evidence that might be pertinent to research questions, among many other lines of evidence, but the reviewers would decide that on a study-by-study basis. If some journal editors regard NHST as virtually never having any value (e.g., Trafimow & Marks, 2015), then they could, for methodological reasons, implicitly or explicitly reject manuscripts reporting use of it. But most important, obtaining an NHST p value less than an arbitrary cutoff would no longer be a necessary precondition for publication, as neither p values *nor any other result* for that matter would be considered in the decision to publish. (I would also think that there should be no bias against publishing good quality replication or exploratory studies or studies that attempt to refute seemingly established theories.)

This RBME editorial policy would be made explicit and prominently stated as a part of journals' guidelines for authors so that investigators considering manuscript submission would be fully aware of it and thus not mistakenly censor or inhibit their own submissions based on what the results of their studies were or were not, because they think a lack of positive or statistically significant findings precludes publication. Any manuscript reviewers, whether in-house or external to the journal, would be fully informed about this policy and implement it. Stated journal submission guidelines for authors should also be explicit and detailed regarding what the journal considers as good methodology.

A two-stage evaluation

When I say “results blind *evaluation*” of submitted manuscripts, I do not necessarily mean that reviewers and editors should literally not look at the Results section of papers in deciding whether to publish. I do not recommend a “results blind *review*” of manuscripts, though some have made recommendations along those lines, such as having authors initially submit *only* the Introduction and Methods sections of manuscripts (Newcombe, 1987). Reported results sometimes indicate some aspects of methodology and data analysis techniques that could not be clearly stated in the Methods/Data Analysis section. Further, Results and Discussion sections certainly have to be edited and reviewed prior to the manuscript going to print, for presentation as well as for substantive reasons. Data analysis in the Results section has to be checked, and correct interpretation of results in the Discussion section has to be assessed as well as a statement of limitations of the study. By “results blind evaluation,” I merely mean that the nature of the observed findings of a study, or whether the study found any effects at all, should be given little or no weight and have no direct bearing on whether the manuscript is to be accepted for publication. The decision for or against publication should be based exclusively on other aspects of the report, especially those that bear on the importance of research questions addressed and the quality of the study's methodology.

As a method of practically implementing RBME, I would recommend a *two-stage* procedure as follows: Authors would submit a manuscript in its entirety to a journal, just as they normally do now. However, upon receipt of it, the editor would distribute it to reviewers after having stripped it of everything except the Introduction and Methods sections

(references and occasional tables or figures referred to in the Introduction and Methods sections would also be left in), to obtain a “first-stage evaluation” of suitability for publication. If the decision of reviewers at this stage is to reject, the authors would be informed that their manuscript was rejected and why. However, manuscripts passing this first stage would then have their abstract and Results and Discussion sections reattached by the editor and submitted to reviewers for a “second-stage review.” A manuscript surviving the first stage as acceptable or judged as requiring no more than minor revisions would be considered “conditionally accepted” for publication, that is, conditional on the absence of any new evidence of flawed methodology discovered in the second stage sufficiently grievous to override the first-stage judgment. Note that the second-stage review merely serves a disconfirmatory or veto function, but where still no weight is given to what the reported results are per se, in the decision to publish. After completion of the second stage, a formal decision is finalized for acceptance, acceptance with revision, or rejection (in the unlikely event rejection is decided at the second stage), and if not rejected, recommendations for minor revision, editing, wording, and cosmetic alterations, and so on, would be suggested or made as pertinent. Such a process would arguably involve about the same amount of person-hours of work as is currently the case, and possibly less, given that reviewers will sometimes have to review only half a manuscript if the first-stage evaluation results in rejection. If it is accepted during the first stage, and the same reviewers who conducted the first-stage evaluation perform the second evaluation, there is no more additional workload than had they reviewed the entire manuscript at once from the onset.

Benefits

With RBME, publication will no longer be influenced by results but decided on, or at least contested on, the playing field of methodology, as it should be. That RBME will reduce publication bias to a large degree is essentially self-evident. There would be no bias at the journal because reviewers cannot decide acceptance based on results if they do not know what the results were. Further, there would no self-censoring bias among investigators because, being fully aware of the RBME journal policy, they would presumably have no inhibition in reporting the null findings of a well-conducted study, knowing that the null finding will in no way reduce their chances for publication. (The absence of an effect found in a well-conducted, important study is no less a reportable finding than the presence of one, and would provide far more valuable information for the scientific community than a reported observed effect of dubious validity and trivial importance that can't be replicated, obtained from a poorly conducted study with insufficient sample size and suspected to be possibly the result of conscious or unconscious machinations to get a significant result required for publication.)

Furthermore, indirectly, many of the endlessly criticized and lamented problems with NHST will be mitigated or made nonissues. Although NHST would not necessarily be banned, authors would feel no compulsion to employ that particular methodology, or any other for that matter, beyond what they deem as most relevant, fitting, appropriate, and methodologically sound for their study. Controversy will no doubt continue over whether NHST has some limited utility in science given proper circumstances of relevancy, and correct application, implementation, and interpretation; however, the widely criticized

overuse, misuse, and overemphasis of NHST p value cutoffs as the predominant gatekeeper of communication of scientific results will end. Most important, publication will no longer be based on effect size at all, whether that effect size is established allegedly inappropriately via NHST, or validly or invalidly *by any other method*. The decision to publish and widely disseminate a study's reported finding will now hinge on whether the reported effect is likely to be true, not what its size or nature is claimed to be (whether large, small, zero, or even counter-intuitive). The only measure that the scientific method provides us of the likelihood of a reported finding being true is the degree to which the methodology of the study reporting it appears sound, independent of the size of any claimed effects or lack thereof.

The problem of irreproducibility of results will also presumably be mitigated as well because there will likely be a reduction in published false positives given that positivity of findings can no longer influence any decision to publish. Null findings judged to be of equal validity to any positive results because of their equally sound methodology will have the same chance of being published as the positive results. Thus, the scientific literature will convey a representative, balanced sample of findings containing a proportion of negative and positive findings duly reflective of what is actually likely to be true. One might say that scientific journals will become more like media outlets with high journalistic standards, reporting important stories that are reliable, confirmed, and based on good sources, rather than being more like tabloids that publish much more sensational reports, but of dubious validity.

With a change in focus of what will determine acceptance of a manuscript for publication, I believe investigators will be forced to not only avoid researching trivial topics but also put greater effort into developing a sound methodology for their study, which would include fashioning clear, explicitly articulated research questions and/or hypotheses, and a stated experimental/study design and subject selection appropriate to addressing those questions, as well as providing a well-thought-out data analysis plan and detailed write-up of the same in a Data Analysis subsection of the Methods section of the manuscript. Further, investigators will have to make a greater effort to ensure that their sample size is adequate, and preferably calculate and formally justify sample sizes up-front in the Methods section where it ought to be, knowing that acceptance of their paper greatly hinges on their having done this. Statisticians and methodological consultants and/or collaborators, as needed by investigators, will be sought earlier in the course of a research project, as they should be, to aid in devising an experimental design and methodology properly suited to research questions, instead of their being called upon late in the game, at the last minute, or after a first round of journal reviews are already made, to do the best they can to address statistical criticisms, "patch up" flawed methodology with ad hoc data analysis techniques, and try to answer vague questions with poor data already collected from a badly designed study with an inadequate sample. In short, investigators will be forced to think more carefully about the substantive issues and research questions their study addresses, as well as the quality of the study's methodology, rather than just rush to get some results, since they know that generally *only* what is in the Introduction and Methods sections of their manuscript will be looked at by reviewers in deciding whether to publish their study report; results will be ignored. As far as publish-ability is concerned, it won't matter what they found in their study or whether they found any "effects" at all. Moreover, any incentive toward deliberate fraud

or unconscious biases in data analysis in order to show positive findings formerly thought to be required for publication would be reduced.

Limitations/criticisms

I cite here some criticisms I have received regarding my suggestion of results-blind manuscript evaluation and my responses to them:

Preregistering studies accomplishes the same thing—Preregistration of studies with funding sources or journals is akin to having papers submitted without their results having been obtained yet, as a way of avoiding any publication bias, precensoring of undesirable findings, or alteration of methods based on results. Although submitting manuscripts for studies that are still only at the planning stage would have the benefit of providing greater assurance that there will be no unconscious publication bias based on results, such policies may sometimes be cumbersome or infeasible to always employ. Study evaluations at the planning stage of a study also allow for earlier reviewer input into the conduct of the study. But preregistration of studies would require additional labor from authors, journal editors, and reviewers, and it is probably not practical and feasible as a widely adopted publication practice. A stated policy of results blind evaluation of entire manuscripts reporting already completed studies would have the benefit of making a required change in journal policies and procedures only among the journal editors and reviewers, whereas authors would seamlessly continue to submit manuscripts just as they did before. The only change is at the journal. (Some journals could conceivably consider a kind of compromise between RBME and preregistering studies by allowing authors to submit only Introduction and Methods sections of a paper before complete execution of the study for provisional evaluation conditional on their following through on their proposed methods as stated.)

RBME would require additional work—No additional work is required by authors as just noted. Only slightly more work is required by journal editors in that they would send a submitted manuscript to reviewers in two waves. No extra work is required by reviewers. In fact, less work would generally be involved for them, as the second half of manuscripts judged unacceptable at Stage 1 would not have to be reviewed, and for most journals the greater proportion of submitted manuscripts are not accepted and therefore would likely fall into this category.

Results are sometimes mentioned in the title or introduction of a manuscript
—The journal's guidelines to authors would have to clearly indicate not to do this, and if authors violate this requirement, the editor can fix it if only trivial editing is needed or return the paper to the authors to change to conform to guidelines just as would be the case for any other violation of submission requirements.

What reviewers regard as important or relevant work is partly subjective and also subject to bias—RBME is not expected to be a panacea for all bias. Some forms will remain, for example, apart from whether findings were positive, authors can decide to submit papers or not, or undertake research projects for that matter, based on various

personal nonscientific reasons. Reviewers have their biases as to what is good methodology. However, methodological preferences must be based to some degree on advocated, consensually validated, rigorous practices, which are not very malleable according to subjective whims. In any case, the impossibility of achieving absolute perfection is no excuse for not doing something that is possible and would produce a net improvement over current practices.

Justifying sample size with formal power analysis presupposes NHST, which may be judged not appropriate methodologically—Some justification for adequacy of sample size would be considered good methodology for most studies, but formal NHST-based power analysis is not the only method for doing this. Precision of interval estimation can be the target for sample size computations. Also, see Trafimow and MacDonald (2017) for an approach to calculating sample size to ensure confidence that estimated effect sizes are close to population values but that is not based on NHST.

Good proposed methodology is not a sufficient criterion for publication: Correct execution and interpretation, which are indicated in Results and Discussion sections, are important too—These issues would be addressed by reviewers in the second stage of review. In the unlikely event of the authors not willing to make relatively minor revisions along these lines and/or not providing adequate justification for what they did, in an otherwise good paper, publication would have to be declined at Stage 2.

Findings reported in the Results section may have a bearing on publication—Not every situation can be foreseen; for example, a novel or unexpected, important finding may have been made, or an unusually strong effect found, that has a bearing on publication, or a medical emergency may dictate reporting of a finding. I'm not suggesting rigorous rules with no flexibility for exceptions. Authors can always make a case for weighing the importance of a finding in their cover letter to the editor. Situations such as these would have to be judged on a case-by-case basis. Still, one must always wonder about the wisdom of publishing any finding based on dubious methodology.

What the study results were has a bearing on publication because findings of small or near zero effects are not important, not worth reporting, and/or are evidence of methodological weaknesses—Whether it was important to test for a given effect should have been argued in the Introduction section of the paper. If it was considered important to test for according to the author and reviewer, then if a small or near zero effect was found, we want to know that. The absence of an effect is just as important to know about as the presence of the effect, assuming reports of both seem truthful. A report of the presence of an effect, if of very questionable truth, should not be promulgated about. Otherwise we have only biased, slanted, cherry-picked information available to us, pretending that it is a truthful, balanced representation of reality. Communicating answers to important questions selectively dependent on what the answers are, is lying by omission.

Summary and conclusions

I suggest that RBME not only solves some of the problems of publication bias and irreproducibility of results, but as a by-product would simultaneously remove the unjustified prominence given to the NHST and arbitrary p value cutoffs in being a gatekeeper of what is reported to the scientific community because now *none* of the results of a study would have any direct bearing on publication, including the NHST p value. Thus, the NHST would in time, it is hoped, become accepted as simply another statistical technique on equal standing with others, which, like any other, can be misused, but if relevant, computed, and interpreted correctly could arguably provide some limited supportive information, but not necessarily more or less than any other method provides. Even if RBME were widely adopted, investigators would at first, by force of familiar habit, no doubt continue to use the NHST, but they will at least now be aware that its result is entirely irrelevant to whether their report gets published. It will no longer be the *results* of the NHST in the Results section of the report, but the *justification for its use* in the Methods section that will be magnified in importance. Although the *BASP* ban on p values may have done statistical science a great service in raising awareness of NHST problems, whether such a ban is necessary as a long-term, explicit, or implicit policy would be a decision of the editor and reviewers of each particular journal based on methodological considerations. Banning p values perhaps would not have been felt necessary if there had not been an exaggerated, unjustified overemphasis of their practical importance to begin with.

There has been decades of controversy concerning criticisms of the NHST. As noted, much of these criticisms are valid, although some are technical and not readily accessible to nonstatisticians. However, it seems to me that most of these problems, as well as publication bias and the so-called crisis of reproducibility and replication, would be greatly ameliorated if we stopped publishing articles on the basis of what was claimed to have been found but rather mostly on the basis of the methodological quality of the study. And I believe the most reliable way to do that is to adopt some sort of situation-suitable variation of a general policy of RBME (at least as a first-stage evaluation).

Even if implementation of proposals made here is found to be impractical for various reasons, it is hoped that these suggestions will at least provoke further thought about workable approximations to them or other ideas that finally address problems related to publication bias, the NHST, and irreproducibility in research.

RBME policies of some kind would, one hopes, contribute to the publication of studies of substantive value and high methodological quality that report key discovered effects as well as important null findings, giving a more valid, globally unbiased, and balanced representation of what is actually true. Impartial information and unbiased evidence for and against theories will thus gradually accumulate and theories will be induced and established in stages of likelihood, conducive to the steady advancement of science. Publication bias of a sort will remain, but it will be of a positive kind. It will shift from a partly unconscious partiality based on results, which often hides and distorts the truth, to conscious selection criteria based on substantive importance and methodological rigor, which more likely reveals it. If journals adopt a policy of RBME, researchers will, for purposes of getting a study

published, have to put their focus and efforts on making sure it is a well-conducted study on an important topic. What they report as findings per se is irrelevant to that purpose, though in the long run the findings of such studies should better approximate important truths. The primary publication bias remaining will be that favoring good studies, and what makes for a good study has nothing to do with what it claims to have found, but rather what it studied, and how it studied it.

Acknowledgments

Thanks to David Trafimow, Ph.D., Department of Psychology, New Mexico State University, and Andrew D. Althouse, Ph.D., Supervisor of Statistical Projects, UPMC Heart & Vascular Institute, Presbyterian Hospital, Pittsburgh, PA, for stimulating discussions and suggestions regarding issues presented in this paper.

References

- Armstrong JS. Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics*. 1997; 3:63–84. DOI: 10.1007/s11948-9970017-3
- Citkowitz M, Vevea JL. A parsimonious weight function for modeling publication bias. *Psychological Methods*. 2017; 22(1):28–41. DOI: 10.1037/met0000119 [PubMed: 28252998]
- Coburn KM, Vevea JL. Publication bias as a function of study characteristics. *Psychological Methods*. 2015; 20(3):310–330. DOI: 10.1037/met0000046 [PubMed: 26348731]
- Cohen J. The world is round ($p < .05$). *American Psychologist*. 1994; 49:997–1003. DOI: 10.1037//0003-066x.49.12.997
- Colom F, Vieta E. The need for publishing the silent evidence from negative trials. *Acta Psychiatrica Scandinavica*. 2011; 123(2):91–94. DOI: 10.1111/j.1600-0447.2010.01650.x
- DeCoster J, Sparks EA, Sparks JC, Sparks GG, Sparks CW. Opportunistic biases: Their origins, effects, and an integrated solution. *American Psychologist*. 2015; 70(6):499–514. DOI: 10.1037/a0039191 [PubMed: 26348333]
- Fanelli D. Do pressures to publish increase scientists' bias? An empirical support from US States data. *PLoS ONE*. 2010; 5(4):e10271.doi: 10.1371/journal.pone.0010271 [PubMed: 20422014]
- Glymour M, Kawachi I. A proposal for editors that may help reduce publication bias. *Letter. BMJ*. 2005; 331:638.
- Greve W, Broder A, Erdfelder E. Results-blind peer reviews and editorial decisions: A missing pillar of scientific culture. *European Psychologist*. 2013; 18(4):286–294. DOI: 10.1027/1016-9040/a000144
- Hanson, R. Conclusion-blind review [Weblog post]. *Overcoming Bias*. 2007. Jan 16. Retrieved from <http://www.overcomingbias.com/2007/01/conclusionblind.html>
- Hanson, R. Result blind review [Weblog post]. *Overcoming Bias*. 2010. Nov 6. Retrieved from <http://www.overcomingbias.com/2010/11/results-blind-peer-review.html>
- Hunter JE. Needed: A ban on the significance test. *Psychological Science*. 1997; 8(1):3–7. DOI: 10.1111/j.1467-9280.1997.tb00534.x
- Ioannidis JPA. Why most published research findings are false. *PLoS Medicine*. 2005; 2(8):696–701.
- Johnson VE, Payne RD, Wang T, Asher A, Soutrik M. On the reproducibility of psychological science. *Journal of the American Statistical Association*. 2017; 112(517):1–10. DOI: 10.1080/01621459.2016.1240079
- Kühberger A, Fritz A, Scherndl T. Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE*. 2014; 9(9):e105825.doi: 10.1371/journal.pone.0105825 [PubMed: 25192357]
- Kupfersmid J. Improving what is published. *American Psychologist*. 1988; 43(8):635–642. DOI: 10.1037//0003-066x.43.8.635
- Locascio JJ. Significance tests and “results-blindness”. *American Psychological Association (APA) Monitor*. 1999; 1999:11.
- Locascio JJ. A statistical suggestion. *Scientific American Magazine, Letters*. 2011 Dec; 305(6):8–10.

- Locascio JJ. June) Psychology and statistics: Readers respond to the BASP ban on p-values. *Significance Magazine*. 2015; 12(3):35–36.
- Maxwell SE, Lau MY, Howard GS. Is psychology suffering from a replication crisis? What does ‘failure to replicate’ really mean? *American Psychologist*. 2015; 70(6):487–498. DOI: 10.1037/a0039400 [PubMed: 26348332]
- Newcombe RG. Towards a reduction in publication bias. *British Medical Journal*. 1987; 295(12):656–659. DOI: 10.1136/bmj.295.6599.656 [PubMed: 3117278]
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015; 349(6251):943.
- Rosenthal R. The ‘file drawer problem’ and tolerance for null results. *Psychological Bulletin*. 1979; 86(3):638–641. DOI: 10.1037//0033-2909.86.3.638
- Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*. 1996; 1(2):115–129. DOI: 10.1037//1082-989x.1.2.115
- Schwarzkopf, S. Is publication bias actually a good thing? (No). *NeuroNeurotic website*. 2015. Retrieved from <https://neuroneurotic.net/2015/10/01/is-publication-bias-actually-a-good-thing/>
- Seife, C. Statistical significance. In: Brockman, J., editor. *This idea must die: Scientific theories that are blocking progress*. New York, NY: HarperCollins Perennial. The Edge Foundation, Inc; 2015. p. 519-522.
- Smulders YM. A two-step manuscript submission process can reduce publication bias. *Journal of Clinical Epidemiology*. 2013; 66(9):946–947. DOI: 10.1016/j.jclinepi.2013.03.023 [PubMed: 23845183]
- Trafimow D. Editorial. *Basic and Applied Social Psychology*. 2014; 36(1):1–2. DOI: 10.1080/01973533.2014.865505
- Trafimow D, Earp BD. Null hypothesis significance testing and Type I error: The domain problem. *New Ideas in Psychology*. 2017; 45:19–27. DOI: 10.1016/j.newideapsych.2017.01.002
- Trafimow D, MacDonald JA. Performing inferential statistics prior to data collection. *Educational and Psychological Measurement*. 2017; 77(2):204–219. DOI: 10.1177/0013164416659745
- Trafimow D, Marks M. Editorial. *Basic and Applied Social Psychology*. 2015; 37(1):1–2. DOI: 10.1080/01973533.2016.1141030
- Van Assen MALM, Van Aert RCM, Nuijten MB, Wicherts JM. Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS ONE*. 2014; 9(1):e84896.doi: 10.1371/journal.pone.0084896 [PubMed: 24465448]
- Van Assen MALM, Van Aert RCM, Wicherts JM. Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*. 2015; 20(3):293–309. DOI: 10.1037/met0000025 [PubMed: 25401773]
- Wang MC, Bushman BJ. Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods*. 1998; 3(1):46–54. DOI: 10.1037//1082989x.3.1.46
- Wasserstein RL, Lazar NA. The ASA’s statement on p-values: Context, process, and purpose. *The American Statistician*. 2016; 70(2):129–133. DOI: 10.1080/00031305.2016.1154108
- Winerman L. How much of the psychology literature is wrong? *Monitor on Psychology*. 2016; 47(6): 14–15.