Psychology of Aesthetics, Creativity, and the Arts
2014, Vol. 8, No. 1, 2–7

1931-3896/14/$12.00 DOI: 10.1037/a0035803

COMMENT

The Empirical March: Making Science Better at Self-Correction

Matthew C. Makel
Duke University

Psychology has been criticized recently for a range of research quality issues. The current article organizes these problems around the actions of the individual researcher and the existing norms of the field. Proposed solutions align the incentives of all those involved in the research process. I recommend moving away from a focus on statistical significance to one of statistical power, renewing an emphasis on prediction and the pre-registration of hypotheses, changing the timing and method of peer-review, and increasing the rate at which replications are conducted and published. These strategies seek to unify incentives toward increased methodological and statistical rigor to more effectively and efficiently reduce bias and error.

Keywords: methodology, power, replication, null hypothesis significance testing, peer review

> "Eighty-seven percent of facts on the Internet are false."
>
> —Sherlock Holmes

The opening quotation is obviously a fabrication. Determining which statements are false is not always so easy, but it is an essential ingredient in how science, as Carl Sagan (1997) explained, winnows deep truths from deep nonsense. In many pivotal places, the social sciences have strayed from this fundamental goal (e.g., Asendorpf et al., 2013; Fanelli, 2010, 2012; Fiedler, Kutzner, & Krueger, 2012; John, Loewenstein, & Prelec, 2012; Nosek, Spies, & Motyl, 2012; Schimmack, 2012; Simmons, Nelson, & Simonsohn, 2011). The current article synthesizes many of the identified ongoing problems and examines some recently proposed strategies to overcome such deficiencies.

One strategy that is receiving substantial attention is replication (e.g., special issues of *The Psychologist*, 2012, and Pashler & Wagenmakers, 2012). Replication is the duplication of research procedures to corroborate previous results (Lykken, 1968; Schmidt, 2009). Schmidt (2009) lists five specific functions of replication: to control for sampling error, to control for artifacts, to control for fraud, to generalize to different and/or larger populations, or to assess the general hypothesis of a previous study. Not all replications serve all functions: Direct replications serve the first and third functions by repeating the original procedures as closely as possible; conceptual replications vary some part of the original procedures (e.g., using a different measure) while still testing the original hypothesis. Conceptual replication may substantiate whether the original findings can be generalized to other contexts. Unlike direct replications, a "failed" conceptual replication does not necessarily refute the original findings; rather, failed conceptual replication may reveal a boundary condition of the original findings or result from poor methods in either the original or replicating studies. Both "successful" and "failed" replications can help reduce false-positive findings, error (both measurement and random), and fraud. Steiger (1990) went so far as to coin the adage, "An ounce of replication is worth a ton of inferential statistics" (p. 176).

Replication is only one of many ways to improve the marketplace of scientific ideas. Replication is good for the field as an ingredient of strong science, but it can be bad for the individual (in the current system) because it is denigrated as bricklaying (e.g., Neuliep & Crandall, 1993). The propulsion model of creative contributions (Sternberg, Kaufman, & Pretz, 2002) similarly casts replication as a "paradigm-preserving contribution" (p. 15) that does not change a field, but rather helps strengthen its current state. Moreover, replication is an inefficient way to resolve concerns about poor research quality because it is, by nature, a follow-up procedure. The best way to avoid bad work is to prevent it from occurring in the first place. How to prevent poor quality research from being disseminated is discussed in further detail later in the article.

The utility of replication grows worse when the numerous biases against conducting and publishing replications are taken into account (Makel & Plucker, 2013a). For example, previous research has shown that in the social sciences, reviewers and journal editors are generally not supportive of publishing replications because they were viewed as a waste of space, a waste of resources, and as providing little information (Madden, Easley, & Dunn, 1995; Neuliep & Crandall, 1993). However, as articulated in the propul-

I thank Jonathan Plucker for the wonderful opportunity to contribute to this special issue and David V. Foster for his helpful comments on a draft of this article.

Correspondence concerning this article should be addressed to Matthew C. Makel, Duke University Talent Identification Program, 1121 W. Main Street, Durham, NC 27701. E-mail: mmakel@tip.duke.edu

sion model, replication is not intended to propel a field forward, but to help assure that the field is actually where it believes to be.

The strength of the bias against replication is clearly illustrated by the relatively low rate at which replications are published in the social sciences. An analysis of the complete publication history of the 100 psychology journals with the highest 5-year impact factors revealed that just over 1% of articles were replications, although that rate has nearly doubled in recent years (Makel, Plucker, & Hegarty, 2012). A similar analysis of the top 100 education journals reported that a dismal 0.13% of articles were replications (Makel & Plucker, 2013a). A third study (Makel & Plucker, 2013b) focused on six giftedness and creativity journals to assess whether domain-specific journals publish replications at higher rates; their replication rate was just over 0.5%.

One ray of sunshine is that all three analyses reported that the majority successfully replicated the original findings. But replications with unique authors were less likely to successfully replicate the original findings than replications in which there was author overlap with the original study. Further, an even darker problem exists in the file-drawer problem (Rosenthal, 1979), in which an unknown number of failed replication attempts could have been rejected by (or never submitted to) journals.

Bias against replication has led many (e.g., Ioannidis, 2012; Makel & Plucker, 2013a; Nosek et al., 2012) to state that there is an overemphasis on novelty in the social sciences. Being wrong is no longer the worst outcome; being unoriginal is worse. Those who were wrong were at least published; those who were unoriginal were never heard from. Emphasizing novel publications breeds a creative process focused almost entirely on idea generation to the exclusion of idea evaluation. Perhaps the epitome of an ill-defined problem, scientific research has myriad goals, multiple approaches to achieving those goals, and numerous viable solutions.

Some may dismiss concern about novelty because science "self-corrects." However, in the education and the giftedness and creativity studies, the median lag between original article and its replication was 7 years (Makel & Plucker, 2013a, 2013b), and in psychology it was 9 years (Makel et al., 2012). Moreover, these long lags only capture the time to potentially identify an error. The time spent discussing and coming to a consensus is obviously much longer. These long lags likely do not satisfactorily keep up with society's desire to implement the knowledge gained from research. Such vacuums give ground to making decisions using criteria other than scientific evidence (e.g., political or self-serving goals). And with a lack of scientific consensus, questionable results will be used to help support political and self-serving goals (e.g., decisions on teacher pay or assessing student progress).

Moreover, when findings actually fail to be replicated, the "correction" does not always become the scientific norm. For example, when presenting on replication at a conference recently, I was surprised to learn that the seminal research on the Pygmalion Effect, showing the power that teacher expectations can play on subsequent student performance (Rosenthal & Jacobson, 1968), had been called into question almost immediately upon publication (e.g., Elashoff & Snow, 1970; Snow, 1995). The original finding has been included in almost every general psychology text I read from high school through graduate school, but I have no memory of reading about the questioning of the findings. Even worse is when publications are fraudulent (for an extreme example, consider *Lancet* retracting a study that falsely connected vaccinations with autism). If immediate questions of validity are not well known, what hope is there for overturning entrenched findings? Have we truly evolved into an anti-Popperian system of research findings being true until proven false (Ferguson & Heene, 2012)?

Fortunately, many have proposed strategies beyond replication to help improve the current state of social science research. As with most complex problems, a whole quiver of interventions is needed to make the necessary improvements. In this article, I group the problems into two levels: the individual researcher and the norms of the field. A third section synthesizes many of the strategies that have been proposed to help the social sciences become better at self-correcting. In aggregate, these strategies seek to unify the incentives of all actors toward increased methodological and statistical rigor, more effectively and efficiently reducing bias and error, both of which combine to make science self-correct more quickly.

Need to Improve the Self-Correction of the Social Sciences

The Individual Researcher

In articles I have been asked to review, there is a trend of simply listing research questions rather than making predictions in manuscripts. The lack of predictions is unfortunate because it blurs the line between predictive and exploratory research. Moreover, when predictions are stated, it appears that social scientists are extremely prescient. In fact, social scientists appear better at predicting outcomes than researchers from any other field (Fanelli, 2010, 2012; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995), including those more deeply grounded in theory and with more rigorous or agreed-upon methods (e.g., space science, geosciences, agriculture, and physics). One possible cause is the use of questionable research practices (QRP; e.g., John et al., 2012; Simmons et al., 2011), such as hypothesizing after results are known (HARKing; Kerr, 1998).

HARKing makes for aesthetically pleasing manuscripts (Giner-Sorolla, 2012; Kerr, 1998) because it assures that all predictions align with all results. Unfortunately, HARKing is but one of many QRPs rampant in the social sciences. A recent survey of psychologists found extremely high rates of self-reports of failing to report all measures, data peeking (collecting additional data if initial analysis dissatisfies), and selectively reporting only results that align with predictions, among others (John et al., 2012). The ability to frame articles in this way has been labeled "researcher degrees of freedom" (Simmons et al., 2011). These practices can be quite subtle, even unconscious. By manipulating methods, analyses, predictions, and how they are reported, researchers are able to present neat and clean results that conform to stories they want to tell. If Lake Wobegon is where all the women are strong, all the men are good looking, and all the children are above average, QRPs mutate research into manuscripts in which all the statistical power is strong, all the p values are good looking, and all the effect sizes are above average.

The Norms of the Field

Researchers also play a role in the work of their peers. Perhaps the two most common and influential roles are serving as review-

ers and journal editors. As reviewers and editors, we accept underpowered manuscripts (Cohen, 1992; Schimmack, 2012). Statistical power is the probability of rejecting the null hypothesis when the null hypothesis is actually false. Statistical power is important because we want to reject the null hypothesis when a finding exists. Additionally, consistently underpowered studies in the social sciences have led to concerns about the prevalence of false-negative findings (e.g., Fiedler et al., 2012).

With pervasive problems such as underpowered studies, not to mention the file-drawer problem and decline effect, false negatives may actually be a more important and more prevalent problem in the field than false positives. Because, as Fiedler et al. (2012)state, "The truncation of research on a valid hypothesis is more damaging and less reversible than the replication of research on a wrong hypothesis" (p. 663). Moreover, false negatives are harder to identify because nonsignificant findings are disappearing from published findings (Fanelli, 2012). We do not want to be the miner who misses a rich vein of gold, but as reviewers and editors, we encourage our peers to put down their pickaxes too soon.

We have known the limits of null hypothesis tests since before almost every currently publishing researcher was a professional (e.g., Bakan, 1966; Rozeboom, 1960). The fallibility of null hypothesis testing was most clearly demonstrated by Bennett, Baird, Miller, and Wolford (2009), when they reported (tongue-in-cheek) statistically significant fMRI data on change in brain activity of a dead salmon. This brings to mind the famous adage, "If you torture the data enough, nature will always confess" (Coase, 1994, p. 27); by using standard statistical techniques, social scientists wrung a confession from a dead fish.

Editors and reviewers making researchers use more conservative alpha levels or corrections for multiple comparisons do not necessarily help avoid false negatives or researcher degrees of freedom. In fact, they could give false confidence to the findings. Many statisticians and methodologists question whether such corrections are appropriate or effective (e.g., Gelman, Hill, & Yajima, 2012). Such contentions suggest that the social sciences have a long way to go before converging on universal reporting methods.

Unfortunately, even when consensus agreements are made about reporting methods, they are not necessarily followed or enforced. For example, the American Psychological Association (APA) "mandated" the reporting of effect sizes in 1999 (Wilkinson & Task Force on Statistical Inference, 1999). Nevertheless, subsequent reports have shown that reporting effect sizes is far from universal (e.g., Giner-Sorolla, 2012; Sun, Pan, & Wang, 2010). The failure of the current system to enforce rigor is further exemplified by a report that found that 15% of published peer-reviewed articles in psychology journals made statistical inference errors (Bakker & Wicherts, 2011). Clearly the peer review/editor system is not functioning as well as it should.

E Pluribus Unum: Aligning Researcher, Journal, and Field Incentives and Goals

In some ways, the Latin phrase for "out of many, one" may seem inappropriate, because this section includes suggestions for transforming the current, uniform, social science research process into diverging paths. However, the underlying goal is to unify the disparate incentives that motivate researchers, reviewers, editors, and journals. In some cases the suggested strategies will make the jobs of individual researchers substantially more difficult, but doing easy work is not the primary goal of researchers.

The following sections introduce five facets of change that help reveal deep truths while reducing bias and error: focusing on facts, not statistical significance; making public a priori predictions; changing both the review and publication processes; and performing replications. All of these facets have been previously suggested in one form or another (e.g., Asendorpf et al., 2013; Nosek & Bar-Anan, 2012; Nosek et al., 2012); my goal here is to introduce them to a new readership and to illustrate how many avenues can be pursued in parallel.

Moving from focus on "significance" to facts. Studies should be accepted or rejected based on how confidently they can answer their question (statistical power), not based on whether they found a question whose answer was "yes" (statistical significance). We know such a filter can exist because the current peer review system has practically eliminated null result publication by rejecting nearly all null results. This statistical significance filter should be replaced with one for statistical power. It is tautological; once manuscripts start being rejected for lack of power, underpowered studies will halt. Moreover, removing the statistical significance barrier would help reveal what Ferguson and Heene (2012) dub "undead theories" that remain alive via fluidity and flexibility, because failed replications would have the possibility for publication.

Following existing mandates (e.g., Wilkinson & Task Force on Statistical Inference, 1999) of reporting and interpreting effect sizes would also help combat power problems (e.g., Paul & Plucker, 2004; Plucker, 1997), while also serving as a stepping stone toward moving beyond directional predictions (e.g., Group A will produce more than Group B) toward making strong predictions about the size of relationships (e.g., Meehl, 1967, 1990; Tukey, 1969).

Putting the "pre" back in prediction. Key stakeholders (e.g., journals, funders, institutional review boards) should begin requiring the preregistration of hypotheses and methods as well as a priori power analyses for predictive research. This action would reduce researcher incentive (and freedom) to perform many QRPs. In fact, it would put the avoidance of QRPs at the focal point of attention, thus making rigor *de rigueur* in the social sciences. Exploratory research also has great value, but needs to be accurately labeled as such, with exploratory research epitomizing the type of results requiring replication.

More radically, acceptance for publication could also be moved to this stage. It already has in some cases (e.g., Association for Psychological Science, n.d.). Instead of researchers HARKing by cleaning, polishing, and shining data through QRPs, and then submitting, revising, editing, and finally publishing, the system could be altered into predicting, preregistering, and then "PARKing": publishing after results known. All the rest can be written and planned a priori. Such preregistration would also help avoid concerns about null findings because acceptance can be determined prior to results being known.

Review and publication. Regardless of whether publication acceptance is moved prior to data collection, both the review and publication processes would be greatly improved if they each were split into two-step processes. Reviewers remain an essential step to the research process and grow even more important in their role as filters of quality as more research is produced. A central clearing-

house of reviewers that is not associated with a particular journal could help make the process of sorting quality more efficient by creating a system that facilitates manuscripts being passed "up" or "down" based on their quality rather than the current "accept" or "reject" system. The separation of publication from peer review would also allow multiple journals to rely on a single set of peer reviews when deciding whether to accept a manuscript. Rubric (http://www.rubriq.com/) and Peerage of Science (http://www.peerageofscience.org/) are examples of peer-review systems unaffiliated with journals that already exist. By removing reviewer allegiance from journals, reviewers can be put in groups based entirely on their areas of expertise.

If editors are worried about how to attract reviewers if they are not attached to specific journals, one possibility is a more formalized pay-as-you-go method, allowing authors to have their articles reviewed if they serve as reviewers for other manuscripts (Fox & Petchey, 2010). If the overlap between quality reviewers and quality researchers is too small, recognizing and rewarding good reviewers in professional evaluation could be of benefit. For example, reviewing could be considered akin to teaching performance. Prolific researchers are often able to buy out of their teaching time; comparable possibilities could be developed for reviewing.

A centralized review system could dramatically reduce the lag time between manuscript completion and actual publication, as this would eliminate the need for researchers to balance which tier journal they submit manuscripts to and how quickly their work is published (assuming submitting directly to a lower tier journal would lead to faster publication). By more quickly placing manuscripts with optimal (in the sense that it would never be "too low" a tiered journal) publication outlets, results would be disseminated more quickly, helping both the individual researcher (earlier opportunity to begin accumulating citations) as well as the field (spreading the new knowledge more quickly).

Similarly, where articles are published could also be determined via centralized clearinghouses. A form of this is already done in publishing law review articles (although, in this format, the review process has not been centralized). Authors submit manuscripts to a clearinghouse site (http://law.bepress.com/expresso/) and select the journals they want to consider their submission. Journals then have an opportunity to review the submission and note whether the manuscript could be published there. Authors may then select where it will be published from the list of accepting journals.

By following in the footsteps of other scientific fields, posting manuscripts to an online archive prior to the review process can also become standard practice. Math and physics have used such a clearinghouse, called arXiv (arxiv.org) for decades. Home to almost 900,000 articles, this open-access site allows authors to post manuscripts prior to the review process and publication in journals. Preposting allows authors to stake claim (and precedence) for research findings without eliminating subsequent publication in a traditional peer-reviewed journal. Social science clearinghouses (e.g., ssrn.com) have existed for decades but have not become the norm for the field. Preposting on such sites removes the review barrier from sharing your work with the world. This is particularly useful when one of the reviewers may have perverse incentives to slow the publication of your work. If preposting became the norm, editors could be charged with scouring prepublication posts to create recommended reading lists (Chopin, Gelman, Mengersen, & Robert, 2013). Similar to posting articles, making data openly available online would also help reduce QRPs because others could independently check analyses. Recent reports in genetics research have shown that posting data is associated with increased citation rates (Piwowar & Vision, 2013).

Replication as it should be. Once the above changes are in place, replication will be free to play the role it was designed to play: corroborating previous results. None of the aforementioned solutions are substitutes for replication, and replication is a poor substitute for all of the above steps at ensuring high quality research.

Determining which studies should be replicated is an important balancing act between maintaining individual intellectual freedom and harnessing resources to facilitate better understanding of constructs of interest. The special section of *Psychology of Aesthetics, Creativity, and the Arts* in Volume 5, Issue 4 (2011), debating the Torrance Tests of Creative Thinking, serves as an example for how "importance" can be determined (especially for ideas that have had time to be more fully evaluated by a broad audience). How was this (and any special issue) topic selected? Key stakeholders identified an area that contained data and disagreement, and determined that a discussion between knowledgeable parties would help advance and clarify understanding and identify potentially fruitful future research paths to help resolve current disagreements.

A similar decision process could be used to help identify what needs to be replicated. Some (e.g., Pashler & Harris, 2012) have advocated for direct replications being conducted prior to conceptual replications, because failed conceptual replications do not necessarily imply uncertainty in the original findings. However, there need not be a centralized decision-making body; individual researchers could make the decision, but so could journal editors or research organizations like APA or American Education Research Association. Lists of pressing issues that need to be resolved could be created; similar to funding agencies releasing requests for proposals on particular topics, journals could release lists of replications for publication in areas of interest. There is precedent for editor (or editorial board) selected content leading to successful output. For example, the journal *Psychological Science in the Public Interest*, founded in 2000, seeks to provide impartial syntheses of available evidence in editorial board-selected areas of national interest (Ceci & Bjork, 2000). As of September 2013, the median number of citations of 28 target articles in the journal's first 10 volumes (2000 to 2009) is 111. Although not replications, this illustrates that editor-driven content need not be low impact.

If citations remain a concern of editors, space could be set aside for the replication of findings from that journal. The space need not be large, because most replications can be extremely brief. Such a policy would not only remind readers of previous findings but also provide an opportunity to strengthen the original findings. Alternatively, findings published in other journals, but that have garnered substantial attention, could also be of potential interest. Koole and Lakens (2012) proposed adopting a method of cocitation in which citing both the original and replicating articles becomes standard practice. This would continue to reward those who got there first, without giving the death penalty to those who did not.

Current initiatives like the Reproducibility Project (Open Science Collaboration, 2012), which is systematically conducting direct replications of articles from top journals, as well as http://

psychfiledrawer.org, a warehouse of replication attempts in psychology, both facilitate an evolution beyond a preoccupation with novelty. Other initiatives to foster replication attempts that have been proposed include incorporating replication into undergraduate and graduate training (e.g., Frank & Saxe, 2012; Neuliep & Crandall, 1993) as well as federal funding devoted specifically to conduct replications (Williams, 2012). All such strategies would help replication serve its role in making the social sciences better at self-correcting.

Discussion

The current article discusses numerous proposals to improve the quality of research being conducted in the social sciences. These proposed strategies include focusing on facts, not statistical significance; making public a priori predictions; changing both the review and publication processes; and conducting replications. Many of these proposals have been around for decades, but have yet to be universally implemented, while others are just beginning to be discussed. As with many interventions, how to scale up these strategies remains to be seen. Doing so will require numerous stakeholders to take the risk of shaking up the status quo.

Working in relatively niche fields, psychologists studying creativity, the aesthetics, and arts often find themselves fighting for attention, resources, and even credibility in the eyes of policymakers. Without a clear track record of high-quality evidence supporting our conclusions and recommendations, our fields may not be able to withstand the type of scandals that have besieged the medical or social psychological research communities when highly cited research fails to replicate (e.g., Begley & Ellis, 2012; Pashler & Wagenmakers, 2012). Therefore, it is better to do our part to ensure that high-quality replicable research is the norm in our fields.

The movements discussed here will not solve all the world's problems or assure our field's prominence. But they should hopefully serve the essential duty of science: incrementally advancing our understanding of the world and helping uncover a host of new, previously unnoticed problems. For the most part, all of the aforementioned strategies are empirical questions waiting testing. If not effective, they should be replaced with alternatives that may be more beneficial. Some (e.g., Nosek et al., 2012) are skeptical of the effectiveness of many proposed solutions, believing that data, materials, and workflow openness is the ultimate solution. Although many interventions that change incentives may not be sufficient solutions, they may be necessary to facilitate such openness. Moreover, open data that is not matched with augmented incentive to check the work of others will lead to little improvement of the current system. There must also be decreased emphasis on novelty and an increased emphasis on correctness.

The empirical march is how we progress toward better understanding how the world works. Our duty as scientists is to strive to increase the tempo of our march by decreasing bias and error and increasing understanding about generalizability and boundary conditions of constructs. Good work is reproducible and strong fields produce reproducible work. Just as science is a process, so is the evolution of the scientific process. To continue progressing, social science needs to transition away from merely discussing the importance of conducting reproducible work to actually producing (and then reproducing) such work.

References

Asendorpf, J. B., Conner, M., de Fruyt, F., de Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. doi:10.1002/per.1919

Association for Psychological Science. (n.d.). *Registered replication reports*. Retrieved from http://www.psychologicalscience.org/index.php/replication

Bakan, D. (1966). Test of significance in psychological research. *Psychological Bulletin, 66,* 423–437. doi:10.1037/h0020412

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43,* 666–678. doi:10.3758/s13428-011-0089-5

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483,* 531–533. doi:10.1038/483531a

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: An argument for multiple comparison correction. Presented at the Human Brain Mapping Conference, San Francisco, CA. Retrieved from http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf

Ceci, S. J., & Bjork, R. A. (2000). Psychological science in the public interest: The case for juried analyses. *Psychological Science, 11,* 177–178. doi:10.1111/1467-9280.00237

Chopin, N., Gelman, A., Mengersen, K. L., & Robert, C. P. (2013). *In praise of the referee*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/referee_v9.pdf

Coase, R. (1994). *Essays on economics and economists*. Chicago, IL: University of Chicago Press. doi:10.7208/chicago/9780226051345.001.0001

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. doi:10.1037/0033-2909.112.1.155

Elashoff, J. D., & Snow, R. E. (1970). *A case study in statistical inference: Reconsideration of the Rosenthal-Jacobson data on teacher expectancy* (Tech. Rep. No. 15). Stanford, CA: Stanford Center for Research and Development in Teaching, Stanford University.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One, 5,* e10068. doi:10.1371/journal.pone.0010068

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

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7,* 555–561. doi:10.1177/1745691612459059

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7,* 661–669. doi:10.1177/1745691612462587

Fox, J., & Petchey, O. L. (2010). Pubcreds: Fixing the peer review process by "privatizing" the reviewer commons. *Bulletin of the Ecological Society of America, 91,* 325–333. doi:10.1890/0012-9623-91.3.325

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science, 7,* 600–604. doi:10.1177/1745691612460686

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5,* 189–211. doi:10.1080/19345747.2011.618213

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7,* 562–571. doi:10.1177/1745691612457576

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7,* 645–654. doi:10.1177/1745691612464056

John, L. K., Loewenstein, G. F., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science, 23,* 524–532. doi:10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217. doi:10.1207/s15327957pspr0203_4

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7,* 608–614. doi:10.1177/1745691612462586

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159. doi:10.1037/h0026141

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How journal editors view replication research. *Journal of Advertising, 24,* 77–87. doi:10.1080/00913367.1995.10673490

Makel, M. C., & Plucker, J. A. (2013a). *Facts are more important than novelty: Replication in the education sciences*. Manuscript submitted for publication.

Makel, M. C., & Plucker, J. A. (2013b, April). *Replications in giftedness and creativity research*. Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7,* 537–542. doi:10.1177/1745691612460688

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–118. doi:10.1086/288135

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141. doi:10.1207/s15327965pli0102_1

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior & Personality, 8,* 21–29.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23,* 217–243. doi:10.1080/1047840X.2012.692215

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615–631. doi:10.1177/1745691612459058

Open Science Collaboration. (2012). An open, large-scale collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7,* 657–660.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536. doi:10.1177/1745691612463401

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. doi:10.1177/1745691612465253

Pashler, H., & Wagenmakers, E. J. (Eds.). (2012). Replicability in psychological science: A crisis in confidence? *Perspectives on Psychological Science, 7,* 528–530.

Paul, K. M., & Plucker, J. A. (2004). Two steps forward, one step back: Effect size reporting in gifted education research from 1995–2000. *Roeper Review: A Journal on Gifted Education, 26,* 68–72. doi:10.1080/02783190409554244

Perspectives on Psychological Science, 7. Special Section on replicability in psychological science: A crisis of confidence? Retrieved from http://pps.sagepub.com/content/7/6.toc

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ, 1,* e175. doi:10.7717/peerj.175

Plucker, J. A. (1997). Debunking the myth of the "highly significant" result: Effect sizes in gifted education research. *Roeper Review: A Journal on Gifted Education, 20,* 122–126. doi:10.1080/02783199709553873

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–641. doi:10.1037/0033-2909.86.3.638

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York, NY: Holt, Rinehart and Winston.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428. doi:10.1037/h0042040

Sagan, C. (1997). *The demon-haunted world: Science as a candle in the dark*. New York, NY: Ballantine Books.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17,* 551–566. doi:10.1037/a0029487

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13,* 90–100. doi:10.1037/a0015108

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. doi:10.1177/0956797611417632

Snow, R. E. (1995). Pygmalion and intelligence? *Current Directions in Psychological Science, 4,* 169–171. doi:10.1111/1467-8721.ep10772605

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180. doi:10.1207/s15327906mbr2502_4

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance: Or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice-versa. *American Statistician, 49,* 108–112.

Sternberg, R. J., Kaufman, J. C., & Pretz, J. E. (2002). *The creativity conundrum: A propulsion model of kinds of creative contributions*. New York, NY: Psychology Press.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102,* 989–1004. doi:10.1037/a0019507

The Psychologist, 25. Special Issue. Retrieved from http://www.thepsychologist.org.uk/archive/archive_home.cfm?volumeID=25&editionID=213

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24,* 83–91. doi:10.1037/h0027108

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. doi:10.1037/0003-066X.54.8.594

Williams, S. N. (2012). Replication initiative: Prioritize publication. *Science, 336,* 801–802. doi:10.1126/science.336.6083.801-c

Received January 6, 2014
Accepted January 6, 2014 ■