Let's Put Our Money Where Our Mouth Is: If Authors Are to Change Their Ways, Reviewers (and Editors) Must Change With Them

Jon K. Maner

Perspectives on Psychological Science 2014 9: 343
DOI: 10.1177/1745691614528215

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2014, Vol. 9(3) 343–351

DOI: 10.1177/1745691614528215
pps.sagepub.com
SAGE

Let's Put Our Money Where Our Mouth Is: If Authors Are to Change Their Ways, Reviewers (and Editors) Must Change With Them

Jon K. Maner
Florida State University

Abstract

A number of scholars recently have argued for fundamental changes in the way psychological scientists conduct and report research. The behavior of researchers is influenced partially by incentive structures built into the manuscript evaluation system, and change in researcher practices will necessitate a change in the way journal reviewers evaluate manuscripts. This article outlines specific recommendations for reviewers that are designed to facilitate open data reporting and to encourage researchers to disseminate the most generative and replicable studies. These recommendations include changing the way reviewers respond to imperfections in empirical data, focusing less on individual tests of statistical significance and more on meta-analyses, being more open to null findings and failures to replicate previous research, and attending carefully to the theoretical contribution of a manuscript in addition to its methodological rigor. The article also calls for greater training and guidance for reviewers so that they can evaluate research in a manner that encourages open reporting and ultimately strengthens our science.

Keywords

methodology, replication, manuscript evaluation

Recently, several scholars in our field have called for fundamental changes in the way researchers conduct and report their work (e.g., Eich, 2014; John, Loewenstein, & Prelec, 2012; Nelson, Simmons, & Simonsohn, 2012; Simmons, Nelson, & Simonsohn, 2011; Simonsohn, 2012; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). These calls are very valuable. Problems with the way researchers operate can be corrected, but for that to occur, we must first engage in open dialogue about what we as a field see as the most pressing problems and the most productive solutions. Communicating about potentially problematic research practices helps the field arrive at consensus viewpoints and encourages researchers to be more open with the way they conduct and report their science.

The goal of the current article is to provide a bit more balance to what I see as an imbalanced aspect of the current discussion over research practices. In our field's treatment of problematic research practices, the responsibility for change—and thus the vast majority of recommendations aimed at improving our system—has been placed almost entirely on researchers (Murayama, Pekrun, & Fiedler, 2013; Simmons et al., 2011; see also Giner-Sorolla, 2012). However, lasting improvements to our science will not be successful unless we also change the behavior of reviewers.

Although others have acknowledged the role that reviewers play in promoting productive research practices, that acknowledgement has consisted primarily of encouraging reviewers to police the system in order to reduce questionable research practices (Nosek, Spies, & Motyl, 2012; Simmons et al., 2011). Little attention has been devoted to changing the general approach reviewers use when evaluating manuscripts.[1] To redress this gap, the current article focuses on the way research practices are shaped by the dynamics among researchers, reviewers, and editors.[2]

Editorial Dynamics

If there are problems with psychological science, the responsibility lies only partially with researchers. In

Corresponding Author:
Jon K. Maner, Department of Psychology, Florida State University, Tallahassee, FL 32306
E-mail: maner@psy.fsu.edu

publishing their work, researchers respond to a set of implicit and explicit criteria used by reviewers to evaluate empirical research. Those criteria, and the general set of scientific expectations they reflect, have helped set the stage for the apparent prevalence of questionable research practices. The behavior of reviewers and editors provides the basis for a powerful incentive system. Researchers often respond to that incentive system by bringing their manuscripts in line with what they perceive as the expectations of those who will be evaluating them. A culture shift in research practices requires a corresponding shift in the manuscript evaluation system—one that is marked by greater acceptance of imperfections in patterns of data and, at the same time, greater attention to overall scientific rigor and theoretical importance.

As reviewers (and editors), we need to begin placing less emphasis on arbitrary and potentially questionable evaluative standards (e.g., Is p less than .05? Are findings perfectly consistent across studies?) and more emphasis on meaningful evaluative dimensions that track overall methodological rigor and potential scientific impact. For example, although psychological data can be—and, indeed, should be—messy, the evaluation process at many top journals rewards "perfection" and involves an expectation that data be extremely clean, in the sense that all effects are statistically significant, the presence of moderating variables is consistent across studies, and the findings are perfectly consistent with hypotheses and previous research (see Giner-Sorolla, 2012).

Although those expectations are rightly intended to ensure that published findings are clear and interpretable, those expectations also promote the very culture that our field is now trying to combat. Such expectations, when applied in a uniform and unyielding fashion across manuscripts, create a clear incentive structure that encourages authors to smooth over imperfections in their data (e.g., failing to report marginally significant or nonsignificant findings; failing to report all dependent variables). However, as others have suggested, it is precisely those papers reporting data that coincide in an apparently perfect way with hypotheses—not those with inconsistencies, marginal results, or other "wrinkles" in the data—that warrant extra scrutiny (e.g., Francis, 2012; Schimmack, 2012; Simmons et al., 2011). In the following sections, I provide specific recommendations to help reviewers prioritize the most replicable and generative studies (these recommendations are summarized in Table 1).

Embracing Imperfection I: Inconsistencies Across Measures

Many studies include multiple dependent variables, and there is some consensus that researchers ought to report all dependent variables that bear directly on the central hypothesis of a paper (rather than selectively reporting only those that show significant results). As that practice becomes more common, reviewers will undoubtedly find themselves in the position of evaluating papers in which the support for a hypothesis varies considerably across dependent variables. If reviewers and editors operate in a "business as usual" fashion, such papers are likely to find themselves criticized for providing a lack of evidentiary support, despite the potential strengths of those papers.

How should reviewers evaluate papers when the strength of the evidence varies considerably across dependent measures? As a starting point, empirical findings do not need to be consistent across dependent variables for a paper to contribute meaningfully to the literature. As others have said (e.g., Eich, 2014), reviewers can require researchers to report all key dependent variables that bear on their hypotheses, but this ought to come with reassurance that some degree of inconsistency across variables can be expected and, on its own, will not be treated as grounds for rejection.

How should reviewers respond if one of two dependent measures provides support for the authors' hypothesis? What about one of five dependent measures? Providing hard-and-fast rules would be difficult and ultimately unproductive, because the level of inconsistency that should be expected and tolerated ultimately depends on the specifics of the paper being evaluated. Instead, let us consider the kinds of questions that reviewers should ask themselves when they encounter situations of this kind.

One critical question is whether the dependent measures are designed to assess the same construct. For example, consider a study that manipulates the experience of rejection and includes three different measures of state self-esteem as dependent variables, only one of which shows an effect that is consistent with the hypothesis (i.e., a decrease in self-esteem). In such a circumstance, the inconsistency across measures is reason for concern because it calls into question the veracity of the self-esteem effect. In this case, it would be appropriate to approach the findings with caution (e.g., by asking for a replication if one is not already provided). Although supporting evidence from one of three dependent variables might not provide an impressive degree of support, supporting evidence from two of three dependent variables would be stronger. That is, reviewer skepticism undoubtedly should vary with the proportion of dependent variables that support hypotheses. Moreover, reviewers might encourage authors to use methods that capitalize on shared variance among dependent variables designed to measure the same construct (e.g., latent variable modeling).

Now consider, in contrast, contexts in which dependent measures are designed to measure different constructs—for example, a rejection study that includes a

Table 1. Recommended Procedures for Handling Common Evaluative Issues While Reviewing

Evaluative issue	Recommended procedure
Inconsistent support across measures	Assess whether dependent variables reflect a common construct (more problematic) or different constructs (less problematic).
	Allow for some degree of inconsistency, and assess whether a preponderance of evidence supports the hypotheses.
	Urge authors to discuss reasons for inconsistencies across measures, and assess the plausibility of those reasons.
Inconsistent support across studies	Assess whether or not the source of inconsistency is a critical linchpin that ties the paper's findings together.
	Urge authors to discuss reasons for inconsistencies across studies, and assess the plausibility of those reasons.
Inferential hypothesis testing	In multistudy papers, focus on meta-analytic findings rather than on individual significance tests.
	Give full consideration to analyses that include covariates, as long as there is some theoretical justification for including them.
	Allow for one-tailed tests as long as there is a clear directional prediction.
	Assess all key inferential tests vis-à-vis statistical power.
Ambiguities regarding the presence of a priori hypotheses	Urge authors to clarify what aspects of their findings were predicted, and evaluate the evidence in light of the presence versus the absence of hypotheses.
	View unpredicted findings with relatively greater skepticism than predicted findings; insist on replication if unpredicted findings seem questionable or are underpowered.
Evaluating a paper's merits	Devote ample review space to discussing the paper's methodological and theoretical strengths.
	Assess and discuss at length the paper's theoretical innovations and importance.
	Assess and discuss at length the paper's potential for solidifying an already existing literature.
Replications	Insist on clear methodological details.
	Give priority to replication attempts for provocative or new findings, whether those attempts succeed or fail.
	Push authors to reflect on possible explanations for failures to replicate (e.g., potential moderating variables).
	In assessing replications, attend more to confidence intervals than to null-hypothesis tests.

measure of state self-esteem and a measure of aggression. The researchers predicted effects for both constructs (reduced self-esteem and heightened aggression), but an effect was observed for self-esteem and not for aggression. This inconsistency across measures could reflect the fact that the effect of rejection does not generalize across the two constructs; perhaps rejection affects self-esteem without necessarily producing aggression. In that case, the lack of effect on aggression should not reduce confidence in the self-esteem effect, and reviewers must assess whether the existing level of support (e.g., a reliable effect of rejection on self-esteem) provides a valuable enough contribution to the literature to warrant publication. In sum, reviewers should regard inconsistencies across measures designed to assess the same (or highly overlapping) constructs as more problematic than inconsistencies across measures designed to assess different constructs.

It is critical that reviewers encourage authors to reflect carefully on inconsistencies across measures. There are many factors that can result in such inconsistencies. Perhaps some variables were not as reliable or valid as others (see Stanley & Spence, 2014, this issue); perhaps the effect on one dependent measure is moderated, whereas the effect on another measure is not; or perhaps some effects are smaller than others, and differences in statistical power render tests on some variables less conclusive. Reviewers should, in turn, reflect carefully on whether the authors have provided satisfying explanations for any inconsistencies. If they have, the paper is likely to succeed in providing readers with a clear sense of what it does and does not achieve and will prevent the findings from appearing cleaner than they actually are. If the authors have not provided satisfying explanations, it would be appropriate for reviewers to question the replicability of the findings.

Embracing Imperfection II: Inconsistencies Across Studies

Many journals in psychology publish multistudy papers, which have the advantage of providing immediate evidence for replicability. Yet with each additional study comes additional opportunities for inconsistencies to emerge. Reviewers sometimes expect multistudy papers to present data that are perfectly consistent across studies. This is not only unrealistic, it is also dangerous, because it encourages authors to hide inconsistencies. For example, authors might be inclined to omit entire studies when those studies do not perfectly replicate patterns of data observed in other studies, thus contributing to publication bias. Such practices provide the field with a misleading portrait of psychological phenomena, hamper meta-analytic efforts to advance a cumulative understanding of a topic, and cause people to form a picture of the findings as clearer and more consistent than they really are. Indeed, as other authors have noted, reviewers should expect inconsistencies across studies, and the presence of such inconsistencies should increase rather than decrease confidence in the findings (Francis, 2012; Schimmack, 2012).

What are reviewers to do when faced with manuscripts that report findings that vary across studies? Such inconsistencies, on their own, should not be grounds for rejection unless those inconsistencies cast doubt on the main story the authors are trying to tell with their data. For example, consider a paper with three studies in which an individual-difference variable is found to moderate an effect in one study but not in the other two. In such a circumstance, researchers might be tempted to ignore the moderating variable in the one study (i.e., fail to report that it was assessed), or they might not report the study at all. My recommendation, in this circumstance, would be for authors to report the moderating variable in all studies and acknowledge that the findings regarding moderation are suggestive but not definitive. My recommendation for reviewers would be to consider carefully whether the paper contributes to the literature without definitive evidence for moderation; if it does, the presence of inconsistencies should not prevent the paper from being published.

Reviewers ought to focus on this question: Is the moderator the linchpin that ties the paper together, or is it a more peripheral aspect of the manuscript? If it provides a central basis for the paper, inconsistent evidence for moderation undermines the paper's contribution, and reviewers should respond accordingly. However, if the effect being moderated (i.e., the main effect) receives consistent support and, on its own, contributes to the literature in some valuable way, then the lack of consistent support for moderation should not detract from evaluation of the paper. In the case of a main effect that sometimes is moderated but sometimes is not, reviewers should caution authors against making too much of the moderator and should encourage them to focus on the main effect. The moderator can be framed as a potentially interesting but inconclusive aspect of the findings that could be pursued in future research. Reviewers should also urge authors to provide plausible explanations for the inconsistency in moderation across studies. Those explanations could involve substantive differences about the methods used in the two studies, or they could involve sampling error or a lack of power to test the interactive effect. Either way, reflecting on any inconsistencies or other imperfections in the data arms readers with a clear sense of what the data achieve and enables readers to more effectively build on the research in their own work.

Inferential Hypothesis Testing

In a single-study paper, individual inferential tests of statistical significance are important, because they provide readers with evidence for the replicability of the phenomenon. However, in a multistudy paper, the replicability of a phenomenon can be assessed without placing ultimate importance on individual inferential tests. Rather, an overall assessment of the strength of the findings can be ascertained by examining the reliability of all the findings reported in the paper. For that reason, not every finding needs to be significant (i.e., $p < .05$) or even "marginally" significant (i.e., $p < .10$). In multistudy papers, individual tests of statistical significance are not a sine qua non for publication (cf. Eich, 2014).

In the case of multistudy papers, reviewers would be wise to insist on internal meta-analyses. Consider the following scenario. A researcher runs four experiments. Two of them produce significant effects. One produces a marginal trend. The other is nonsignificant. In these circumstances, the researcher should include all four studies, ideally showing in a meta-analysis that the finding is reliable across studies (see also Braver, Thoemmes, & Rosenthal, 2014, this issue). However, in evaluating such a paper, many reviewers might criticize the lack of consistent evidence for statistical significance and point to the nonsignificant findings as a lack of evidentiary support. But this would be a mistake. The overall weight of the evidence ought to be considered, and the strength of the evidence should not rest on individual tests of statistical significance but instead on a meta-analysis of a paper's overall findings.

The meta-analytic approach I suggest can be contrasted with a common strategy some researchers currently use, given the culture of the editorial system at our top journals. Some authors might try to publish just the

first two significant studies, perhaps also including the marginal finding in the third, because of the (probably correct) belief that the greater consistency depicted in such a paper would satisfy reviewers' desire for empirical cleanliness. This approach, however, makes the finding appear cleaner and more robust than it really is. It produces a biased estimate of effect size. It reduces the likelihood of replication, because it does not provide researchers the information they need to estimate statistical power or to attend to potentially relevant moderating variables that account for variability in effect size across studies. And it perpetuates a culture in which authors and reviewers conspire to keep empirical imperfection out of our journals. The time has come for a new strategy, one that focuses on meta-analytic findings; embraces a reasonable amount of inconsistency; and provides clear, accurate, and replicable portraits of empirical findings.

Another issue that has received significant attention involves the use of covariates. For example, Simmons et al. (2011) suggested that researchers who include covariates must report what their findings would look like without the covariates. This request is quite reasonable, but how should reviewers respond when findings are weakened by the exclusion of covariates? In such a case, reviewers may be inclined to view the inclusion of covariates as cheating and to reject a pattern of results that fails to hold when covariates are excluded. I think this tendency, when applied in a uniform manner across papers, is a mistake. Because covariates can account for systematic variance in the dependent variable that would otherwise be attributed to error, researchers often measure covariates precisely because they can enhance tests of statistical significance and produce more accurate measures of effect size. This is a perfectly legitimate empirical approach. As long as there is some theoretical basis for including a covariate—and that basis can be as simple as noting the presence of common variance between the dependent variable and the covariate—conclusions based on analyses that include covariates should be given full consideration.

An additional recommendation pertains to the use of one-tailed versus two-tailed significance tests. Historically, two-tailed univariate tests have been the norm in psychology, despite clear reasons for using one-tailed tests when the researchers have a directional hypothesis. Demanding two-tailed tests is inappropriate, especially in multistudy papers in which an initial result is predicted and replicated in a subsequent study. Although one-tailed tests of a directional hypothesis are legitimate, reviewers tend to view them as providing only weak evidence for a hypothesis. Without some impetus, authors are likely to continue using two-tailed tests, and reviewers are likely to continue insisting on them. I therefore encourage reviewers to embrace the use of one-tailed tests when a clear directional hypothesis exists.

This recommendation requires that authors be forthcoming about the presence of directional hypotheses. Saying one predicted a particular finding when one did not is a violation of ethical scientific principles. Other scholars have taken up this issue, and for further insight, readers can consult Bem (2004) and Kerr (1998). In my own view, it is perfectly reasonable to say, after the fact, that some unpredicted finding is consistent with an existing literature and could have been predicted based on that literature. But it is not permissible to say that one actually predicted a finding if one did not. (And only in the case of a priori directional predictions can one-tailed tests be used legitimately.) When the presence of such predictions is ambiguous, reviewers should push authors to clarify whether they truly predicted their results.

Indeed, this point is worth emphasizing again: Reviewers should urge authors to be clear in communicating whether or not a pattern of results was predicted a priori. The way reviewers evaluate a manuscript should hinge, in part, on whether the findings were predicted based on some theoretical framework. Aside from the issue of one-tailed versus two-tailed tests, a pattern of results that was predicted warrants substantially greater confidence than a pattern that was unpredicted (Murayama et al., 2013; Popper, 1963). Findings that were unpredicted should be approached with greater caution, particularly if they have not been replicated. In my view, reviewers do not push authors enough on this point, and authors are sometimes too ready to describe any pattern of results as being predicted. Asking authors to clarify which aspects of their findings were predicted a priori arms reviewers with information that is integral to assessing the strength of the evidence. Moreover, reviewers might weight predictions recorded prior to an experiment as especially compelling (e.g., preregistered predictions or those written and saved in a lab prediction archive ahead of time), given that researchers' memories for their own predictions can be biased by motivated reasoning (see Nosek et al., 2012).

A final point about inferential hypothesis testing: Inferential tests should be assessed vis-à-vis statistical power. It is important that studies use sample sizes that are large enough to provide adequate statistical power (see Lakens & Evers, 2014, this issue). Statistical tests that lack power (i.e., power considerably lower than .80; see Cohen, 1988)—even those that reach conventional levels of statistical significance—can be regarded with skepticism if they have not been replicated and are based on relatively small samples (see Francis, 2012). In contrast, tests that do not reach conventional levels of significance but are in the ballpark (e.g., $p < .15$), are adequately powered, and reflect effect sizes of impressive-enough magnitude to contribute to the literature should be viewed more positively by reviewers. In such circumstances, reviewers can also ask authors to report

confidence intervals around the key parameter estimates to provide readers with information about the possible magnitude of the effect (see Cumming, 2014).

The Editorial System Would Benefit From Greater Promotion Focus

Some might view the editorial process in psychology as being guided a bit too much by a focus on preventing Type I error, as opposed to a focus on promoting interesting and important research findings (cf. Higgins, 1992). That is understandable, given the apparent prevalence of possible Type I errors in our field. Nevertheless, even our best journals sometimes seem more inclined to publish unimportant papers with seemingly incontrovertible data than potentially groundbreaking but imperfect papers. Reviewers should not lose sight of the fact that the editorial system is ultimately designed to disseminate interesting and important research.

It is a truism to say that reviewers should carefully evaluate the interest value and importance of every manuscript. However, many reviewers do not spend enough time evaluating a paper's virtues, in the sense of identifying and highlighting potential contributions to the literature. My sense is that reviewers sometimes view their main job as weeding out the weak—preventing papers with problems from passing too easily through the system—as opposed to carefully evaluating the positive aspects of a manuscript or thoughtfully considering its contribution to the literature. Reviewers should not relegate their discussion of a paper's merits to a few boilerplate sentences in the opening paragraph (If I only had a nickel for every review that said "This paper is interesting, timely, and well-written. However . . . "). Reviewers should devote considerable space (i.e., more than just a few sentences) to reflecting on each paper's merits and discussing its potential contribution to the literature.

There are multiple ways for papers to contribute to the literature. For example, some papers are very provocative, are theoretically innovative, or provide a large conceptual advance that could eventually change the way the field thinks about a particular phenomenon. Such manuscripts can be extremely generative. In my own view, the field would benefit from prioritizing the publication of such papers—despite the notable presence of imperfections—as long as the research is methodologically rigorous and the authors readily acknowledge and discuss any inconsistencies, ambiguities, or other limitations of the data.

In contrast to papers that are very novel or innovative theoretically, some papers provide an incremental advance over existing studies, but they do so in a methodologically rigorous way that solidifies the literature and increases the field's confidence in existing knowledge. Such papers, like those that make major theoretical advances, are valuable to the field and should be given priority in the evaluation process. Lack of novelty or originality is a frequently cited reason for rejection, and this can stifle researchers' ability to publish scholarship that overlaps with previous work. Good science can be incremental, and reviewers should not reject a paper simply because they are aware of previous articles that report similar findings. Indeed, given the importance of replication, manuscripts that report rigorous research involving small deviations from previously published studies can be quite valuable, as long as they help solidify a literature in some important way.

Reviewing Manuscripts Involving Replications

Replication is a critical aspect of science. Consequently, papers that replicate (and, ideally, extend) previous findings should be given full consideration, even at top journals. A worthwhile approach would be to devote some proportion of journal space to replication attempts (see LeBel & Campbell, 2013, for an example). In such cases, reviewers should not require elaborate theoretical details if those details are already available in the literature. Reviewers should, however, insist on clear methodological details so that readers can assess the precision with which the replication was conducted. It is critical that reviewers assess whether the replication attempt is adequately powered and based on sound methods that accurately recreate the methods from previous research (Zwaan, 2014; see also Perugini, Gallucci, & Costantini, 2014, this issue).

Reviewers should be prepared to recommend publication not just for papers that replicate—but also for those that fail to replicate—important findings within the field. Although failures to replicate (and other null findings) can provide valuable insight into some psychological phenomenon, it is not uncommon for reviewers to criticize manuscripts because the results are inconsistent with previous findings or because they break with conventional wisdom in the field. One possibility is that reviewers view a failure to replicate as indicative of some problem with a study's methodology. This, however, incorrectly affords precedent to the original findings being replicated. As already discussed, some published reports may be difficult to replicate because the original reports glossed over imperfections in the data so as to adhere to reviewer expectations. If a finding has been replicated multiple times by different labs, one can be reasonably confident in that finding, and thus, failure to replicate it warrants skepticism. However, relatively new findings or those that have not been extensively replicated should not be considered definitive and treated as

a baseline from which to evaluate new research. In such cases, a failure to replicate can be just as informative as a successful replication.

On encountering a replication attempt, reviewers should consider this key question: In what way would this replication study be important or informative? If the paper focuses on a relatively new or controversial topic, for example, a replication attempt stands to make a valuable contribution, whether that attempt succeeds or fails. Even replication studies involving findings that seem well-established can be useful, particularly when there is some question as to whether the literature has been influenced by publication bias or other negative research practices. In contrast, if a paper replicates a finding that has been independently replicated multiple times by multiple labs, it might not contribute as much to the literature.

Failures to replicate may help identify critical moderating variables, and in those cases, reviewers should urge authors to reflect on variables in the person or situation that might serve as boundary conditions. When confronted with evidence that disconfirms previous findings, reviewers should not discount the paper, but should insist that authors attempt to explain why their findings differ from those in previously published reports. Papers that succeed in doing that should receive priority in the editorial process.

How best to evaluate the success or failure of a replication attempt is still being debated. For example, some have argued that, in evaluating replication papers, confidence intervals are more useful than null-hypothesis testing (see Cumming, 2014). Rather than assessing whether or not the effect being replicated is statistically significant, one should assess whether the confidence interval around the parameter estimate for the effect contains the estimate observed in the study being replicated. In this view, reviewers should attend carefully to confidence intervals when evaluating replications and should use that information (rather than null-hypothesis tests) to evaluate the success or failure of the replication and, ultimately, to assess the value of the manuscript. However, the field has not yet achieved consensus on this issue, and other approaches have been suggested (e.g., using meta-analysis to aggregate the parameter estimates from the original study and replication study). Reviewers should attend to the changing views on this topic, as replication studies are likely to become more and more common.

Closing

Right now is a time of transition for our field. Expectations are changing, as are research practices. There is bound to be some substantial uncertainty, and consequently, authors are likely to stick with whatever research practices they think will satisfy reviewers and result in their work being published. Breaking this vicious cycle will require communication and dialogue among editors, reviewers, and researchers. As the discussion our field is having about research practices begins to settle into areas of consensus, reviewers and editors should communicate clear guidelines, expectations, and reassurances to authors.

There are bound to be growing pains as our manuscript evaluation system evolves. Some researchers may be gun shy, worrying that their work will be readily discounted because of inconsistencies or other limitations of the data. Reviews and action letters should explain to authors that some level of empirical imperfection is to be expected and that imperfection, on its own, does not jeopardize a paper's potential for contributing to the literature. Even better, journals should communicate this point ahead of time to editors, reviewers, and authors by revising reviewer and author guidelines. Equipped with a greater sense of safety, authors may then be more inclined to put all their cards on the table.

Positive change in the manuscript evaluation system would also be facilitated by providing graduate students and other early career professionals with formal training on how to review and evaluate manuscripts. Indeed, I believe manuscript evaluation training should be an integral part of graduate training, and graduate students would benefit from opportunities to write practice reviews and from receiving concrete feedback from senior scholars. Students should be taught to evaluate research in a reasonable and balanced manner: to carefully weigh a paper's imperfections along with its merits and to consider the suggestions articulated above along with others provided by the recent literature (e.g., Simmons et al., 2011).

The recommendations described in the current article are meant to supplement other ideas, such as opening up the research bottleneck to alternative forms of scientific information sharing (e.g., Giner-Sorolla, 2012; Nosek & Bar-Anan, 2012) and increasing the value placed on null results (Ferguson & Heene, 2012). We as a field need to push authors to publish science that is methodologically rigorous and to more readily acknowledge the imperfections in their work, while at the same time encouraging reviewers to be more accepting of those imperfections. To be clear, my recommendations are not meant to lower reviewers' standards but, rather, to shift them. Reviewers should shift their attention away from arbitrary standards (e.g., Is p less than .05?) and toward more meaningful evaluative dimensions (e.g., Does a meta-analysis indicate that the findings are reliable?) and an overall assessment of methodological rigor and potential scientific impact. This shift serves the goal of prioritizing the most generative and replicable studies.

As reviewers, we would also be wise to remember that most authors are themselves also reviewers. Adhering to arbitrary and outdated reviewer standards that demand perfection is likely to perpetuate a vicious cycle, such that authors of the manuscript we are reviewing may then go on to demand perfection when they review other manuscripts (possibly our own). With this point in mind, it would be sagacious to live by these words: Review unto others as you would have others review unto you.[3]

The manuscript evaluation process reflects a constructive collaboration among editors, reviewers, and authors. One concern that arises from the discussion currently unfolding in our field is that it sometimes portrays researchers as perpetrators and reviewers as police. However, it is important to recognize that the relationship among authors, reviewers, and editors is dynamic; each responds to changes in the other. If the expectations of reviewers and editors, and thus the incentive structure of the editorial process, remains the same, authors are unlikely to change the way they conduct and report research. For the culture of our science to change, we need to put our money where our mouth is, and we all need to change together.

Acknowledgments

The author is grateful for constructive comments from Jim McNulty, Mary Gerend, Bo Winegard, Alison Ledgerwood, and Bobbie Spellman.

Declaration of Conflicting Interests

The author declared that he had no conflicts of interest with respect to his authorship or the publication of this article.

Notes

1. Some scholars have proposed dramatic changes to the basic nature of the editorial system (e.g., making the whole review process public; Nosek & Bar-Anan, 2012; Nosek et al., 2012). Commenting on those changes is beyond the scope of the current article. I assume that if such changes do occur, they will be slow to implement, and changes designed to optimize the current editorial system would still remain useful.
2. Little systematic or current data are available on the expectations of editors and reviewers. I therefore base my description of the editorial process on many interactions I have had with authors, editors, and reviewers, as well as my own experience as a researcher, reviewer, and editor for two major journals in social psychology (*Journal of Personality and Social Psychology* and *Personality and Social Psychology Bulletin*). Consequently, my characterization of the editorial process may not apply to all journals, editors, or reviewers in psychology or even to most of them. However, I feel confident in saying that they do apply to at least a sizable subset of those entities.
3. I am grateful to Bobbie Spellman for providing these nice words of wisdom.

References

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

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*, 333–342.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6.

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.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156.

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.

Higgins, E. T. (1992). Increasingly complex but less interesting articles: Scientific progress or regulatory problem? *Personality and Social Psychology Bulletin*, *18*, 489–492.

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.

Lakens, D., & Evers, E. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*, 278–292.

LeBel, E. P., & Campbell, L. (2013). Heightened sensitivity to temperature cues in individuals with high anxious attachment: Real or elusive phenomenon? *Psychological Science*, *24*, 2128–2130.

Murayama, K., Pekrun, R., & Fiedler, K. (2013). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*. Advance online publication. doi:10.1177/1088868313496330

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2012). Let's publish *fewer* papers. *Psychological Inquiry*, *23*, 291–293.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: Opening scientific communication. *Psychological Inquiry*, *23*, 217–243.

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.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*, 319–332.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London, England: Routledge.

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

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. (2012). It does not follow: Evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press). *Perspectives on Psychological Science*, *7*, 597–599.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*, 305–318.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432.

Zwaan, R. A. (2014). Replications should be performed with power and precision: A response to Rommers, Meyer, and Huettig (2013). *Psychological Science*, *25*, 305–307.