Reviewing the Peer Review*

Gaurav Sood
gsood07@gmail.com

December 1, 2015

NB: Preliminary draft. Please do not cite without permission.

Science is a process. And for a good deal of time, peer review has been an essential part of the process. Looked independently, by people with no experience with it, it makes a fair bit of sense. For there is only one well-known way of increasing the quality of an academic paper— additional independent thinking. And who better than engaged, trained colleagues?

But this seemingly sound part of the process is creaking. Today, you can't bring two academics together without them venting their frustration about the broken review system. The plaint is that the current system is a lose-lose-lose—the authors, the editors, and the reviewers, all lose enormous amounts of time. And that there is little to show for the time— gains from implementing suggestions are thought to be generally low, and sometimes negative.

More egregiously, peer reviews aren't able to catch simple, common, important errors. For instance, Acharya, Blackwell and Sen (2015) find that 67% of the relevant set of empirical articles in top political science journals condition on post-treatment variables. Similarly, Nieuwenhuis, Forstmann and Wagenmakers (2011) find that in nearly half of the relevant articles in top cognitive neuroscience journals mistakenly claim that the difference in a significant result and insignificant result is significant (see also, Gelman and Stern, 2006).

The problem that consumers of peer review most often lament, however, is not its incapacity to catch basic methodological errors, but the preponderance of subjective taste based judgments;

*The note benefited from comments by Pablo Barberá, Scott Clifford, Justin Esarey, Andy Guess, Thomas Leeper, Brendan Nyhan, and Daniel Stone.

the judgments are often veiled as criticisms about framing. So strong is the feeling about the thinness of the rationales behind the negative recommendations that *Reviewer 2* has become a meme.[1] Doubtless, subjectivity of the basis of reviews, contributes to the other major complaint—capriciousness of the review process. Many academics think that the draw of the luck plays a much larger role in deciding the fate of the paper than its quality.

The data suggest that there is something to the academics' claims. For instance, a National Award winning book when submitted anew to 27 publishers and agents was rejected by all, including the original publisher of the book (Ross, 1982). But social science isn't fiction, or so Peters and Ceci (1982) thought. They submitted 12 published articles within three years of publication; 8 of the 12 were rejected for "serious methodological flaws." More temporally proximally still, a systematic assessment of peer review during the Neural Information Processing Systems conference revealed that two independent committees charged with deciding on whether or not to accept the manuscripts disagreed about 57% of the accepted papers (Price, 2014).

Reliability of the review process likely has many important consequences. Most obviously, it bears on decisions about the number of reviewers that journals need to reliably accept or reject the same article. But why ought we care about a reliable peer review process? Because reliability of the peer review affects the quality of the journal (accepted article). The point is best illustrated with the aid of an example. Imagine a top ranked journal in the field. Assume also that the journal arbitrarily restricts journal space (see, for e.g., Card and DellaVigna, 2014). And assume that the journal wants to publish top X articles. Assume also that there is a commonly agreed upon metric called 'worthiness' that reviewers measure in their reviews, paying no consideration to any other factors. If reviewers measure perfectly, the true top X articles would be chosen every time, maximizing the "worth" of the published articles (journal). But say the reviewers are not reliable, for e.g., in trying to balance competing demands, some don't read the article closely, making an error in estimating the article's worth. Such errors cannot increase the "worth" of a journal, as it

[1]For instance, "We wish to thank Reviewer 2 for their critical feedback and sincerely apologize for not having written the manuscript they would have written." (Twitter) or "Reviewer 2 walks into a bar and leaves promptly, complaining of it being the worst coffee shop they've ever seen." (Twitter)

cannot be increased. The effect of errors is always zero or negative; errors will have no impact on the quality of published articles or reduce the quality.

The example, however, understates the problem. The "worth" of an article is generally thought to be decided on the basis of assessments of two, conceptually uncorrelated, constituent items— quality and importance (see also, Ellison, 2000). And even if there is consensus on quality, judgments about importance appear to be multi-dimensional, if not arbitrary. Assuming that importance in reality isn't subjective, judgments about importance are an important source of error.[2] Worse, academics suspect that judgments about importance guide judgments about quality, that reviewers reason in a motivated manner.[3] To address this, Leeper (2015) proposes a bifurcated review process, with judgments about quality (research design) made separately from judgments about importance. The system also precludes post hoc reasoning about research design depending on whether or not the results appeal to reviewer's prejudices.

Lower reliability has other potential downstream consequences. Partly because everyone knows that the review process is so noisy, there is an incentive to submit articles that people know aren't good enough. Submitters may think (correctly) that the chance of a 'low quality' article being accepted is reasonably high. Thus, low reliability systems may increase the number of submissions (Baghestanian and Popov, 2014). Increase in submissions, in turn, is likely to lower the quality of the reviews, and lower the reliability of recommendations still further. It is a vicious cycle. And the answer may be as simple as making the process more reliable.

One way to make the process more reliable is to increase the number of reviewers per paper. But that risks increasing the burden on reviewers, which may reduce quality of each review sufficiently to offset gains from larger n. Asking more reviewers to review a paper is but one way to increase the reliability of the review system. Providing reviewers guidance is another. We could, for instance, provide reviewers a brief white paper on common statistical issues to watch out for, including some of the problems noted above— controlling for post-treatment variables

[2]See Esarey (2015) for a discussion of how differences in reviewers' judgments of "quality" can yield systems with lower reliability.

[3]Worries about motivated reasoning apply more directly, and with yet greater force, to reviews of articles that contain results that do not agree with the prejudices of reviewers.

(Acharya, Blackwell and Sen, 2015) and treating difference in significant result and insignificant result as significant (Nieuwenhuis, Forstmann and Wagenmakers, 2011). Or, the editor could include a link to brief notes on important issues with links to more detailed treatments in their invitation to reviewers. More general guidelines for reviewers would be better still.[4] What should reviewers attend to? What are they missing? And perhaps most critically, how do we incentivize this process?

When thinking about incentives, there are three parties whose incentives we need to restructure: the author, the editor, and the reviewer. Authors' incentives can be restructured by making the process less noisy, as we discuss above. Knowing that an article would be reliably rejected from a journal of a particular rank alters the cost (additional time)-benefit (publication in a higher ranked journal) calculations for submitting an article to that journal. Other more conventional costs—time and money—can also be used to manipulate incentives. For instance, a cost—payable in time or money—can be levied for rejection. And perhaps a smaller cost for each Revise and Resubmit.

As for the editors— if the editors are not blinded to the author (and the author knows this), they are likely to factor in the author's status in choosing the reviewers, in whether or not to defer to the reviewers' recommendations, and in making the final call. Thus, we need triple-blinded pipelines. On the other end, whether or not the reviewer's identity is known to the editor also likely affects reviewer's contributions, in both good and bad ways. For instance, there is every chance that junior scholars in trying to impress editors file more negative reviews than they would if they would if they knew that the editor had no way of tying the identity of the reviewer with the review. But keeping the identity of the reviewer from the editor runs the risk of reducing incentives to review.

Another, perhaps more effective, way to incentivize reviewers would be to publish the reviews publicly, perhaps as part of the paper. Just like online appendices, we can have a set of reviews published online with each article. Making reviews public means the incentives for reviews

[4]See Bernstein (2008) and Lucey (2014) for some advice on how to review an article.

become roughly the same as they are for the broader academic market—some premium for quality, which in turn is some interaction between being correct, and original insight. And we already have a few "working" versions of public review systems—blogs and published 'exchanges' between authors and critics. More recently, Publons offers a system that credits reviewers for peer review.

The discussion hitherto has been mostly superficial. Neither the severity of the problems, nor the efficacy of the cures has been established. Good data are needed for both. But reviews aren't available to be analyzed. And given how critical a component peer review is in scientific production, they very likely should be. Minimally, we stand to gain a better understanding of the failings and strengths of the review process. And that may motivate informed experimentation. And very plausibly, making reviews public would have a few other positive externalities, it would incentivize authors to exert yet more effort before they submit the papers for review. Today, the inevitability of a *Revise and Resubmit* plausibly reduces incentives to submit a more finished manuscript.

But when we talk about releasing data, what data do we mean? The text of the reviews, for one. It could be used to quantify the quality of reviews. And figure out all the ways in which we fail. If a census is too much to ask for, a more limited agenda can be pursued. For instance, it would be good to know how often are reviewers the source of bad statistical advice. If there are sensitivities around releasing review text, permissions can be sought from reviewers and authors; consent screens can be incorporated into peer review systems. And we may only release data where everyone involved agrees, or ask everyone involved for their consent to be bound by other decision rules at the start of the process.

If the editors are hesitant, a group of scholars can come together and crowd-source collection of review data. People can deposit their reviews and the associated manuscript in a specific format to a server. And to maintain confidentiality, we can sandbox these data allowing scholars to run a variety of pre-screened scripts on it. Or, journals can institute similar mechanisms.

Aside from full-text of reviews, a variety of meta data about the review process can be

released. For instance,

- Whether a manuscript was desk rejected or not
- How many reviewers were invited
- Time taken by each reviewer to accept (NA for those from whom you never heard)
- Total time in review for each article (till Revise and Resubmit or Reject) (A separate column for each revision.)
- Time taken by each reviewer
- Recommendation by each reviewer
- Length of each review
- How many reviewers did the author(s) suggest?
- How often were suggested reviewers followed-up on?
- School ranking of the author(s), reviewer(s)
- Stage of career of the author, reviewer
- Topical area of the paper (from the menu provided to the authors), and broad category of research $\sim$ theory, methodology, empirical (separate entries for quantitative and qualitative)
- Race and gender of authors

In fact, much of the data submitted in multiple-choice question format can probably be released easily. These data, in turn, promise to shed light on basic, but important questions about the health of the review process. For instance, conditional on certain attributes and reviewer recommendations, do acceptance rates vary by gender?

But capitalizing on 'found data' ought not to define the boundaries of our efforts to study and improve the peer review process. More data that sheds light on peer review should be actively

collected. For instance, journals could randomly allocate the same article to separate independent committees and empirically assess the reliability of judgments. Or, they could run experiments around specific conjectures. For instance, editors of *Journal of Public Economics* experimentally tested whether instituting shorter deadlines for reviewers reduces review times (They do.) (Chetty, Saez and Sándor, 2014). More ambitious experiments, like ones that encourage a set of reviewers and authors to go public would be interesting to pursue.

Of the many areas of research political science contributes to, its contributions to the understanding of institutions are among the most vital. One of the reasons it has been able to make some modest progress on understanding institutions, which are some of the hardest social scientific questions, is because the discipline has long understood, and been comfortable with, co-existence of strategy and irrationality (or boundedness of rationality). Thus, it is uniquely positioned to make progress on reforming the systems that produce science. The question is also vital, for contributions to systems that produce knowledge carry with them the promise of greatly enhancing human welfare.

Here's to making advances in production of knowledge, to pursuit of truth.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2015. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects.".

Baghestanian, Sascha and Sergey V Popov. 2014. "On Publication, Refereeing and Working Hard." *Refereeing and Working Hard (September 25, 2014)* .

Bernstein, Marc. 2008. "Reviewing Conference Papers.".

Card, David and Stefano DellaVigna. 2014. "Page Limits on Economics Articles: Evidence from Two Journals." *The Journal of Economic Perspectives* pp. 149–167.

Chetty, Raj, Emmanuel Saez and László Sándor. 2014. What Policies Increase Prosocial Behavior? An Experiment with Referees at the Journal of Public Economics. Technical report National Bureau of Economic Research.

Ellison, Glenn. 2000. The slowdown of the economics publishing process. Technical report National Bureau of Economic Research.

Esarey, Justin. 2015. "How Does Peer Review Shape Science? A Simulation Study of Editors, Reviewers, and the Scientific Publication Process."
http://jee3.web.rice.edu/peer-review.pdf.

Gelman, Andrew and Hal Stern. 2006. "The difference between âĂIJsignificantâĂİ and âĂIJnot significantâĂİ is not itself statistically significant." *The American Statistician* 60(4):328–331.

Leeper, Thomas. 2015. "Has the Time Come for Bifurcated Peer Review?"
http://thomasleeper.com/2015/06/bifurcated-peer-review/.

Lucey, Brian. 2014. "10 tips from an editor on undertaking academic peer review for journals.".

Nieuwenhuis, Sander, Birte U Forstmann and Eric-Jan Wagenmakers. 2011. "Erroneous analyses of interactions in neuroscience: a problem of significance." *Nature neuroscience* 14(9):1105–1107.

Peters, Douglas P and Stephen J Ceci. 1982. "Peer-review practices of psychological journals: The fate of published articles, submitted again." *Behavioral and Brain Sciences* 5(02):187–195.

Price, Eric. 2014. "The NIPS Experiment." http://blog.mrtz.org/2014/12/15/the-nips-experiment.html.

Ross, Chuck. 1982. "Rejecting published work: Similar fate for fiction." *Behavioral and Brain Sciences* 5(02):236–236.