INQUIRY: The Journal of Health Care
Organization, Provision, and Financing
Volume 59: 1–3
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00469580221090393
journals.sagepub.com/home/inq
SAGE

An Empirical Assessment of Reviewer 2

Christopher Worsham, MD[1,2,3], Jaemin Woo, BA[1], André Zimerman, MD[4], Charles F. Bray, BS[1], and Anupam B. Jena, MD, PhD[1,2,5]

Abstract

According to research lore, the second peer reviewer (Reviewer 2) is believed to rate research manuscripts more harshly than the other reviewers. The purpose of this study was to empirically investigate this common belief. We measured word count, positive phrases, negative phrases, question marks, and use of the word "please" in 2546 open peer reviews of 796 manuscripts published in the British Medical Journal. There was no difference in the content of peer reviews between Reviewer 2 and other reviewers for word count (630 vs 606, respectively, P = .16), negative phrases (8.7 vs 8.4, P = .29), positive phrases (4.2 vs 4.1, P = .10), question marks (4.8 vs 4.6, P = .26), and uses of "please" (1.0 vs 1.0, P = .86). In this study, Reviewer 2 provided reviews of equal sentiment to other reviewers, suggesting that popular beliefs surrounding Reviewer 2 may be unfounded.

Keywords

peer review, reviewer 2, research, publication, journals

KEY POINTS:

What is already known about this topic?

According to research lore, the second peer reviewer (Reviewer 2) is believed to rate research manuscripts more harshly than the other reviewers, yet this has not been empirically investigated.

What does this study add?

This is the first empiric analysis comparing Reviewer 2 to other reviewers in biomedical research.

What are the implications of this study?

Contrary to popular belief, Reviewer 2 may not rate research manuscripts more harshly.

Introduction

The editorial and peer review process for research manuscripts can be challenging for investigators as they submit their work to journals for publication. Despite being considered a fundamental aspect of the dissemination of scientific research, the peer review process is flawed and susceptible to bias.[1,2] For example, a single-blinded peer review process (where reviewers know who the authors are but the authors do not know who the reviewers are) has been associated with preferential publishing of studies from high-profile authors and institutions compared to a double-blinded process, which may make it more difficult for less experienced investigators to publish their work.[3]

One aspect of the peer review process that has attracted academic and popular attention is the level of criticism offered by specific reviewers. In particular, some researchers believe that the second peer reviewer of submitted research manuscripts (fondly referred to as "Reviewer 2") rates the manuscript more harshly than other reviewers, as evidenced by the Facebook group "Reviewer 2 Must Be Stopped!", which has over 76,000 members as of this writing. Although this appears to be a popular perception, empirical evidence to support or refute this hypothesis is scant, suggesting currently held beliefs about Reviewer 2 are largely based on individual anecdotal experiences that may be subject to confirmation bias based on existing legend in a flawed peer review landscape.

Methods

To test differences between Reviewer 2 and other reviewers, we analyzed 2546 initial open peer reviews of 794 research manuscripts published in the British Medical Journal (BMJ) from 2015 to 2020 that were evaluated by 2 to 5 peer reviewers. We focused on peer reviews from manuscripts' first decision because not all manuscripts had subsequent peer reviews. In BMJ decision letters, Reviewer 2 is the second reviewer to return comments, irrespective of the order in which reviewers were asked to evaluate a manuscript. For each review, we tallied the word count, negative phrases, positive phrases, question marks, and use of the word "please," under the assumption that harsher peer reviews would be longer, with more negative phrases, questions, and requests (eg, "please" conduct a given analysis), but fewer positive phrases. We manually classified phrases as negative or positive based on a blinded assessment of an automatically generated list of commonly used phrases in the sample (Supplementary Table 1**)**. To test differences between Reviewer 2 and other reviewers, we estimated a separate linear regression for each of the outcomes (ie, 5 separate regression models with outcomes including: number of words, negative phrases, positive phrases, question marks, and word "please") with a binary variable for Reviewer 2 (key independent variable). We first estimated an unadjusted review-level model, which simply compared means of each of the above outcomes between reviews performed by Reviewer 2 vs other reviewers. Next, we estimated an adjusted model accounting for manuscript-level fixed effects (primary analysis), making the adjusted model a within-manuscript analysis. By including manuscript-level fixed effects into each outcome regression, this approach effectively compared the sentiment of reviews performed by Reviewer 2 with the sentiment of reviews performed by other reviewers *for the same manuscript*. After estimation of each adjusted model, we calculated the adjusted mean outcomes (eg, adjusted number of words in reviewer reports by Reviewer 2 vs other reviewers) using the marginal standardization form of predictive margins.[4]

Results

The average peer review was 612 words, used 8 negative phrases, 4 positive phrases, 5 question marks, and the word "please" once. Of 794 articles, 249 (31.4%) had 2 reviewers, 244 (30.7%) had 3 reviewers, 189 (23.8%) had 4 reviewers, and 112 (14.1%) had 5 reviewers.

For each of the 5 outcomes, there were no significant differences between Reviewer 2 and other reviewers both before and after regression adjustment (Figure 1**;** Supplementary Table 2). The adjusted word count for Reviewer 2 and other reviewers was 630 and 606, respectively (P = .16). The adjusted number of negative phrases was 8.7 for Reviewer 2 and 8.4 for other reviewers (P = .29), and the adjusted number of positive phrases was 4.2 for Reviewer 2 and 4.1 for others (P = .10). The adjusted number of question marks was 4.8 for Reviewer 2 and 4.6 for other reviewers (P = .26). Finally, the adjusted number of instances the word "please" was used was 1.0 for both Reviewer 2 and other reviewers (P = .80).

Discussion

In a text analysis of open peer reviews of published medical research manuscripts, we found that contrary to common belief, there was no difference in sentiment in reviews by Reviewer 2 compared to those of other reviewers. These findings are consistent with a study in political science that was focused on reviewer recommendations and not text analysis.[5] Our study was limited by the consideration of accepted manuscripts at a single journal, which may result in greater concordance across reviews, and at a journal with an open review policy, representing a small fraction of articles submitted to scientific journals.[6] However, open review policies have been shown not to affect review quality and publication recommendation, suggesting that our results may hold even when reviews are not public.[7,8] In the BMJ, Reviewer 2 is simply the second reviewer to return an evaluation; in other journals, this may not be the case, limiting generalizability of this study to journals who use a similar review process. Our study was also limited by a manual determination of key words or phrases that might suggest a review was negative. Another approach would have been to manually characterize all reviews as being positive or negative or alternatively, classify a subset of reviews manually and then train a machine learning based algorithm to predict review sentiment. Finally, the origin of the Reviewer 2 lore is also unclear and may simply reflect a general frustration of authors with the peer review process rather than an issue with any specific reviewer. Reviewer 2, unfortunately, seems to have received the brunt of this frustration, though our findings suggest that is unwarranted.

[1]Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
[2]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA
[3]Division of Pulmonary & Critical Care Medicine, Massachusetts General Hospital, Boston, MA, USA
[4]Postgraduate Program in Cardiology and Cardiovascular Sciences, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
[5]National Bureau of Economic Research, Cambridge, MA, USA

Received 17 July 2021; revised 3 March 2022; revised manuscript accepted 9 March 2022

Corresponding Author:
Anupam B. Jena, MD, PhD, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115, USA.
Email: jena@hcp.med.harvard.edu

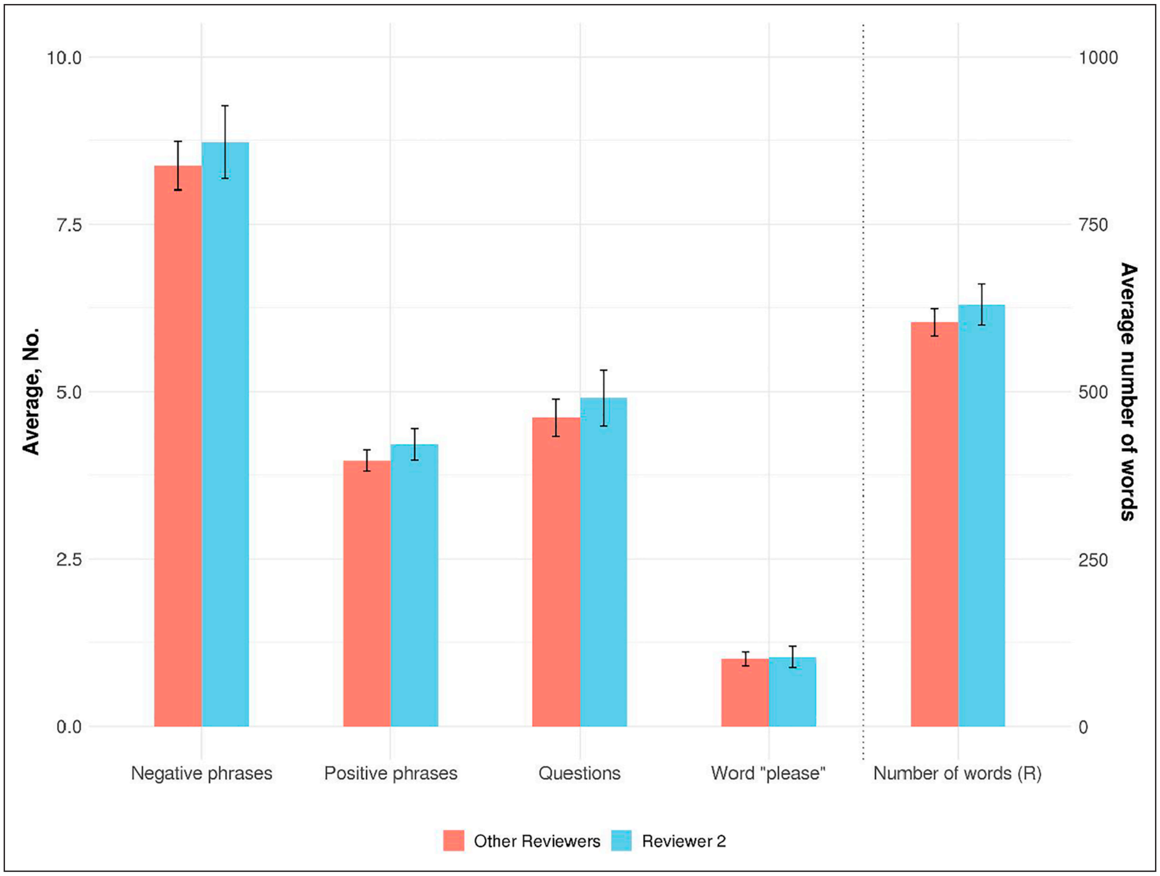

Figure 1. Comparison of reviews performed by Reviewer 2 and other reviewers. Note: Adjusted averages with error bars representing 95% confidence intervals.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Dr. Jena reports receiving (in the last 36 months): consulting fees unrelated to this work from Bioverativ, Merck/Sharp/Dohme, Janssen, Edwards Life Sciences, Novartis, Amgen, Eisai, Otsuka Pharmaceuticals, Vertex Pharmaceuticals, Celgene, Sanofi Aventis, Precision Health Economics, and Analysis Group; income unrelated to this work from hosting the podcast Freakonomics, M.D., and from book rights from Doubleday Books. Dr Worsham reports receiving (in the last 36 months): consulting fees unrelated to this work from Analysis Group, FVC Health, Chronius, NuvoAir, and Alosa Health; income unrelated to this work from book rights from Doubleday Books.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Christopher Worsham https://orcid.org/0000-0002-1611-6871

Supplemental Material

Supplemental material for this article is available online.

References

1. Bohannon J Who's Afraid of Peer Review? *Science* 2013;342(6154):60-65. doi: 10.1126/science.342.6154.60.
2. Haffar S, Bazerbachi F, Murad MH. Peer Review Bias: A Critical Review. *Mayo Clin Proc* 2019;94(4):670-676. doi: 10.1016/j.mayocp.2018.09.004.
3. Tomkins A, Zhang M, Heavlin WD. Reviewer bias in single- versus double-blind peer review. *Proc Natl Acad Sci Unit States Am* 2017;114(48):12708-12713. doi: 10.1073/pnas.1707323114.
4. Williams R. Using the margins command to estimate and interpret adjusted predictions and marginal effects. *STATA J* 2012;12(2):308-331. (In English). doi: 10.1177/1536867x1201200209.
5. Peterson DAM. Dear Reviewer 2: Go F' Yourself. *Soc Sci Q* 2020;n/a(n/a). doi: 10.1111/ssqu.12824.
6. Polka JK, Kiley R, Konforti B, Stern B, Vale RD. Publish peer reviews. *Nature* 2018;560(7720):545-547. doi: 10.1038/d41586-018-06032-w.
7. van Rooyen S, Godlee F, Evans S, Smith R, Black N. Effect of blinding and unmasking on the quality of peer review: a randomized trial. *JAMA* 1998;280(3):234-237. doi: 10.1001/jama.280.3.234.
8. van Rooyen S, Delamothe T, Evans SJ. Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial. *BMJ* 2010;341:c5729. doi: 10.1136/bmj.c5729.