

COMMENTARY

An empirical exploration of reviewers' and editors' roles fostering high quality research during peer review

Roxanne Ross* and Eric D. Heggstad

University of North Carolina at Charlotte

*Corresponding author. Email: Rross37@uncc.edu

Köhler et al. (2020) argue that peer review acts as an essential mechanism to foster high-quality work and as a vetting process that protects the field from low-quality science. Reviewers are to support this task by checking that the research methods employed are rigorous and transparently reported. They are also to offer suggestions to authors to improve their manuscripts in ways that adhere to high standards of research conduct. The focal article describes some concerns about how reviewers are engaging in these tasks. Some of the concerns raised relate to encouragement of questionable research practices (QRPs) during peer review and the prioritization of statistically significant or clean results over methodological rigor.

As it turns out, we rather recently conducted a survey that can provide insight into these two issues and, as such, offers some granularity on the inner workings of peer review in management and applied psychology. More specifically, our data provide insight regarding the extent to which reviewers and editors (a) suggest the use of QRPs; (b) observe and respond when authors, other reviewers, or editors use or suggest using QRPs; and (c) evaluate a manuscript based on patterns of support for the hypotheses. To obtain these data we emailed a total of 3,118 reviewers and editors. Our final sample was 297 respondents, including 140 reviewers and 157 editors (i.e., senior or associate). Reviewers were identified as editorial board members and individuals who had published in top management and applied psychology journals¹ during the five-year period 2012–2016. Senior and associate editors were identified by examining masthead information for 204 management and applied psychology journals. Reviewer and editor samples were combined for ease of presentation as we found no meaningful differences between the two groups in the results to be presented for this commentary. The survey was pre-registered with the Open Science Framework. The preregistration contains the survey material and an anonymized database, which are viewable using this link: https://osf.io/q8cny/?view_only=6e2d6fc405024706845aebcdbcae875a

Do reviewers/editors suggest QRPs?

Previous research informing the argument that reviewers and editors are suggesting authors use QRPs have relied on authors' interpretations of reviewers' and editors' behaviors (Banks et al., 2016; Butler, Delaney, & Spoelstra, 2017; LeBel et al., 2013). Authors' perspectives are important, but it is also important to evaluate the extent to which reviewers and editors recognize and acknowledge that they are asking authors to engage in QRPs. To investigate this issue, we

¹Journals included *Academy of Management Journal*, *Administrative Science Quarterly*, *Journal of Applied Psychology*, *Journal of International Business Studies*, *Journal of Management*, *Journal of Organizational Behavior*, *Leadership Quarterly*, *Organizational Science*, *Personnel Psychology*, and *Strategic Management Journal*.

presented survey respondents with six hypothetical “comments to authors” that we created to be similar to comments that might be provided by a reviewer or editor in the review process. Three of these comments were constructed to serve as an exemplar for when a reviewer or editor encourages the authors to engage in a specific QRP. Comments reflected hypothesizing after results are known (HARKing) (i.e., “a hypothesis should be changed to be more consistent with findings”), selectively reporting hypotheses (i.e., “a hypothesis should be taken out because its results were not statistically significant”), and selectively reporting samples (i.e., “a sample should be removed because it produced results that did not support the hypothesis, were not statistically significant, or were messy”). These QRPs were chosen because they seemed the most pertinent to the reviewing stage. To reduce demand characteristics, the other three comments were written to address similar aspects of the article but that did not encourage the authors to engage in a QRP: (a) suggesting authors take out a hypothesis because it had a weak theoretical justification or was theoretically uninteresting, (b) suggesting the authors remove a sample from the article because the associated data were not related to the main purpose of the manuscript, and (c) a suggestion, because of unexpected findings, to conduct additional analyses and report them as exploratory. For each of these six hypothetical reviewer comments, respondents indicated how frequently they had made similar comments when evaluating a manuscript (response scale: never, rarely, occasionally, often, and almost all of the time).

One-third of the respondents in our sample (34%) indicated they had made a comment when evaluating a manuscript that was similar to at least one of the three QRP-related comments (i.e., a response other than “never”). Across the individual QRPs, 20% of respondents indicated that they, at some point, had made a comment about HARKing, 19% reported making a comment about selectively dropping a hypothesis, and 16% reported that they had made a comment about selectively reporting samples. These results highlight that while most reviewers report never suggesting these three particular QRPs to authors, they also indicate that QRP-related suggestions are being made in the review process.

Respondents who acknowledged suggesting one of the QRPs were asked to describe the reasons why they had offered such a comment. A dominant theme across all three QRPs was that these kinds of comments were symptomatic of other problems with the study. For example, respondents commonly indicated that they would suggest HARKing when they felt that there was another hypothesis that was equally or more applicable and that was more in line with results presented. Similarly, respondents indicated that they had provided authors with comments to remove a hypothesis when problems with the theoretical foundations or limitations in the methods used could explain the lack of support. For example, one reviewer noted, “I recommend this only when there are research design problems that likely contributed to the non-significant result.” The most common reason given for recommending selectively removing a sample was also because of notable theoretical or methodological limitations.

Interestingly, these results differ markedly from those of Banks et al. (2016), who asked authors about receiving QRP-related suggestions from reviewers and editors. Their study found that 40% of authors said that they had been asked to selectively report hypotheses, 33% had been asked to HARK, 14% had been asked to selectively include control variables, and 10% had been asked to exclude data. Misinterpretation of reviewers’ and editors’ comments could potentially explain why our results differ markedly from those of Banks et al. (2016). It is possible that well-intentioned suggestions (e.g., drop a sample because of methodological problems) are sometimes being interpreted as questionable by authors (e.g., drop a sample because the results did not support the hypotheses). Perhaps, in an effort to prevent authors from misinterpreting a comment, reviewers and editors could be encouraged to offer more thorough explanations for their comments that could be perceived by authors as a QRP. Potential misinterpretations could be avoided if reviewers and editors were to explain *why* they were making a particular suggestion.

Are reviewers/editors acting as “gatekeepers”?

Köhler et al. (2020) mention that reviewers and, particularly, editors should direct authors not to engage in QRPs and/or to not follow the advice of reviewers when they suggest a QRP. Our data provide some detail on these points. Namely, we asked whether reviewers and editors were noticing and intervening when authors used QRPs or when others on the review team suggested them. In our survey, we presented respondents with a definition of a QRP provided by Banks et al. (2016) to ensure a standardized understanding of the term. We then asked, “Have you noticed authors using QRPs to improve results or improve statistical significance?” (response options: yes or no). Fifty nine percent responded that they had seen authors engage in QRPs. When participants indicated they had seen such behavior, we asked them how they responded. Reviewers frequently asked authors for additional information or justification for the questionable aspect of the article in their reviews. It was also common for reviewers to suggest authors not engage in the QRP. Editors, as would be expected, tended to be more direct: they typically reported that they would either try to tactfully address the QRP in their decision letter or simply reject the article. For example, one editor mentioned, “I just clearly state that I have noticed that QRPs seem to be present and ask for an explanation or counter explanation.” Another editor was even more direct, stating that he/she had gone so far as to ban authors from submitting additional work to the journal and had written letters to the heads of the author’s department.

We also asked reviewers and editors if they had noticed “other reviewers or editors suggesting authors use QRPs to improve results (e.g., in order to make the results appear less messy) or improve statistical significance?” Thirty percent of our sample indicated they had observed editors/reviewers suggesting QRPs to authors. For those that responded “Yes,” we asked what they would do in that situation (i.e., “Did you respond to what you noticed? Please describe what action you took if any”). Reviewers most typically indicated that it was not their place to do anything. As one reviewer commented, “I give my opinions but I’m not here to tell other reviewers (let alone an editor) what to do.” Editors, in contrast, were likely to ask the authors to ignore the comment from the reviewer and/or to replace the reviewer in the next round of peer review.

These results suggest both some strengths and weaknesses of how reviewers and editors are acting as gatekeepers and protecting the field from problematic research conduct. When authors appeared to use QRPs, many reviewers and editors reported directly discussing their concerns with them. However, protection from other reviewers’ or editors’ suggestions to use QRPs was less robust. Reviewers who had witnessed such practices from other reviewers or editors did not feel empowered to report them, and editors often claimed to use relatively indirect tactics to address these types of comments from reviewers. Although many editors reported communicating with authors, few mentioned that they would directly address a reviewer who made a questionable comment. This result suggests reviewers may not receive much feedback about their behavior. A solution to this problem is to have more communication from the editor. For instance, editors could directly discuss with reviewers that their comments may be construed as a QRP and that those types of comments should be avoided. Without such communication, reviewers may suggest QRPs again or use them in their own work. More communication from editors may help when they decide to reject an article because authors seemed to have used a QRP. Editors could explain that concerns about the use of a QRP was one contributing factor leading to rejection rather than simply rejecting the manuscript without explicitly connecting the decision with the use of a QRP.

Do the results influence the evaluation?

Köhler et al. (2020) contend that the review process may have a bias toward statistically significant findings, such that reviewers and editors may be more likely to recommend accepting an article when statistical analyses show support for hypotheses. Our survey included a series of questions in

which respondents were asked to indicate the extent to which the nature and pattern of the statistical results would likely lead them to feel that an article should be rejected. Specifically, we had them rate the extent to which (a) too many unsupported hypotheses, (b) inconsistent results, and (c) findings that conflicted with previous research would lead them to have a more negative view of the manuscript (i.e., recommending rejection) (response scale: never, rarely, occasionally, often, and almost all of the time). Sixty-three percent of respondents indicated that having too many unsupported hypotheses is a factor that can lead to recommending rejection (i.e., any response except “never”). Similarly, 77% indicated that inconsistent results can lead them to recommending the rejection of a manuscript. Finally, 40% of respondents indicated findings that are inconsistent with prior research can lead to recommending a rejection. Overall, only 18% of respondents indicated they never consider any of these three factors when deciding to reject a manuscript.

The focal article’s claim that reviewers may view the pattern of statistical significance or the direction of the results as a proxy for research quality appears to be reflected in our data. Interestingly, fewer participants stated that findings that conflicted with previous research factored into a rejection decision than for the other two types of results. This finding supports the claim in the focal article (and from past work; Edwards & Berry, 2010; Nosek, Spies, & Motyl, 2012) that reviewers and editors tend to have more favorable attitudes toward novel or counterintuitive findings. Although participants showed less bias toward counterintuitive results, 40% is still a relatively high number.

These explicit preferences for certain types of results point to a need to redefine the norms of the field. It is worth reiterating the focal article’s point that theoretical or practical contributions and methodological rigor are not necessarily the same as statistically significant or clean results. If the premise and execution of a research project are sound, null results should still be informative. Conflating study findings with methodological and theoretical quality potentially incentivizes QRPs, as authors know publishing is a competitive process and are likely to be keenly aware of anything they can do to increase the chances of success. On a more practical note, systematic bias toward certain results makes informative discovery harder. For instance, it may take longer to refute statistically significant findings that have already been published. If counterintuitive findings are suppressed, it may take longer to identify an unknown moderator that can explain the inconsistencies across different studies.

Social learning theory offers some insight as to how norms could be changed to be more open to certain kinds of results. The theory suggests that people emulate the behaviors of others, and they are more likely to emulate those who hold status in prestigious hierarchies and who control rewards (Bandura, 1977). This suggests that editors’ actions are particularly important. The actions of the editors in top journals likely have the strongest influence, as they hold high status and control coveted rewards. Editors could send a strong signal by adopting two-stage reviewing. This type of action would convey that the editor finds it necessary to consider results distinctly from the quality of other aspects of the article. As a result, others may question the belief that certain types of results are less valuable or indicative of low-quality science. Two-stage reviewing also addresses reviewers’ tendency to have a more implicit bias toward null or messy results (e.g., Emerson et al., 2010) by disentangling reviewers’ evaluation of theory and methodology from the results.

References

- Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Banks, G. C., O’Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., . . . Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, 42(1), 5–20. doi: 10.1177/0149206315619011

- Butler, N., Delaney, H., & Spoelstra, S.** (2017). The gray zone: Questionable research practices in the business school. *Academy of Management Learning & Education*, *16*(1). doi: [10.5465/amle.2015.0201](https://doi.org/10.5465/amle.2015.0201)
- Edwards, J. R., & Berry, J. W.** (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods*, *13*(4), 668–689.
- Emerson, G. B., Warme, W. J., Wolf, F. M., Heckman, J. D., Brand, R. A., & Leopold, S. S.** (2010). Testing for the presence of positive-outcome bias in peer review: A randomized controlled trial. *Archives of Internal Medicine*, *170*(21), 1934–1939. doi: [10.1001/archinternmed.2010.406](https://doi.org/10.1001/archinternmed.2010.406)
- Köhler, T., González-Morales, M. G., Banks, G. C., O’Boyle, E., Allen, J. A., Sinha, R., . . . Gulick, L. M. V.** (2020). Supporting robust, rigorous, and reliable reviewing as the cornerstone of our profession: Introducing a competency framework for peer review. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *13*(1), 1–27.
- LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T.** (2013). PsychDisclosure.org grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science*, *8*(4), 424–432. doi: [10.1177/1745691613491437](https://doi.org/10.1177/1745691613491437)
- 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*(6), 615–631.