Industrial and Organizational Psychology (2020), **13**, 72–75
doi:10.1017/iop.2020.8

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

Navigating the review process through the holier than thou

Jeffrey B. Vancouver*

Ohio University
*Corresponding author. Email: vancouve@ohio.edu

As the focal article suggests, reviewing is a complex job that requires sophisticated knowledge and skills. One guiding recommendation included within the integrity competency presented in the focal article (Köhler et al., 2020) mentioned refraining from encouraging questionable research practices like HARKing. HARKing (i.e., hypothesizing after results are known) includes adding, revising, or suppressing hypotheses non-transparently once the results are known (Kerr, 1998). Yet, the constructive competency included giving actionable advice to authors and the representing competency asked reviewers to evaluate the work in terms of the outlet (e.g., a journal) and its constituents in order to vet the article and thus promote "the dissemination of high quality, relevant, and rigorous work products" (Köhler et al., 2020, p. 4). My concern is that some of the behaviors one might categorize as HARKing, and thus would presumably violate the integrity competency, may be needed to fulfill the constructive or representing competencies. For example, suggesting an article be rejected "encourages" the suppressing of one or a set of hypotheses in a way that will not be transparent to the larger scientific community. In this comment, I discuss some of the more nuanced and not so nuanced issues related to reviewing and HARKing, with the idea that some of the lines being drawn regarding prescribed reviewer behaviors might be interfering with the more productive elements of the reviewer's role.

Discouraged HARKing

The not so nuanced issue regarding HARKing begins with the example provided above. Anyone who reviews for journals knows that few articles submitted are worthy of publication. In some cases, this might be because of poor fit with the journal, but often it is because the research is so weak in terms of theoretical rationale and methodological rigor that the information value it might add to the field does not exceed the cost incurred in crowding out more worthy articles. For such low information articles, a reviewer's duty is to recommend rejection to the decision maker and to constructively explain the issues to the authors.

Yet, for some articles, the issues of weak theoretical rationale and poor methods may be true for only *some* of the hypotheses or some aspect of the hypothesis. Many of us have read shaky rationales for hypotheses whose tests are subject to many alternative explanations and/or methods so weak or confounded that one might be more concerned about the disinformation value than the information value of the results reported. This is particularly true when the purpose of examining the hypothesized relationship is to test causality, given that threats to internal validity must be considered in addition to threats to construct, statistical conclusion, and external validity (Shadish, Cook, & Campbell, 2002). In some cases, the flaws are symmetrical such that whatever is found (i.e., significant or non-significant) is suspect, but other times it is asymmetrical (e.g., concern that the task used created floor or ceiling effects, or that the measurement or

© Society for Industrial and Organizational Psychology, Inc. 2020.

https://doi.org/10.1017/iop.2020.8 Published online by Cambridge University Press

manipulation was so weak that null results cannot be reasonably interpreted as disconfirming the theory). Meanwhile, one might see that other aspects of the study are of enough quality, perhaps with some refocusing, that the information value of the article could be constructively honed into a contribution worthy of the audience the authors want to reach. Given the guidelines presented by Köhler et al. (2020), this situation creates a dilemma for the reviewer because it involves suggesting that certain hypotheses, likely ones not supported, be suppressed.

The horns in this dilemma stem from the integrity and the constructive competencies described in the focal article. The integrity horn rests on the assumption that HARKing is bad. Indeed, HARKing *might* result in three ills. First, a hypothesis added or revised after results are known might occur because one developed an explanation or theory from a spurious effect (i.e., "fudging"; Lipton, 2005). Second, suppressed hypothesis tests might bias the population of observed effects meta-analysts include in a review. Third, suppressed hypotheses might increase the possibility that others repeat the same flawed test in the future (i.e., ignorance of history). To negotiate this dilemma, let me provide guidance regarding these ills for reviewers.

The first ill, the development of explanations or theory from spurious effects, is both common and unlikely. It is common because the explanations and theories in our field are nearly all derived inductively or abductively (i.e., they are derived to explain existing observations). That is, hypotheses derived a priori are often just extrapolations of previous findings or grounded "theory" that summarizes previous findings. It is the logic and rationale, perhaps with diagnostic empirical support (i.e., challenging tests), that should provide the a priori likelihood estimate of the verisimilitude for the explanation made by a theory for an effect. Unfortunately, a deep dive into the past literature often reveals that the logic is thin and the empirical tests weak in that they eliminate few alternative explanations for previously observed effects. To be sure, if the effect has been replicated, this removes capitalization on chance as one explanation. Thus, although the effect is not likely spurious, the explanation attached to it may be. Consider Lipton's (2005) example of a fudged theory: epicycles. Epicycles could predict the position of planets quite well (i.e., they were empirically robust and replicable). However, what we now know is that the explanation for epicycles was contrived to fit the geocentric view of the solar system.

It is our duty as expert reviewers to make sure authors acknowledge thin logic and methodological weaknesses when interpreting their own results as well as the literature used to derive their hypotheses. A constructive suggestion for authors is often to highlight the weaknesses in the existing literature in the introduction of a article. This often further justifies the information value of the focal study, assuming the focal study does not repeat the same flaws as the existing research. If the study does repeat the same flaws as the previous literature, the reviewer should probably recommend rejecting the article.

At the same time, developing new explanations for a spurious effect is *unlikely* because there are already so many theories and explanations in existence that finding one for the authors who did not seem to have found it prior to the study is often not difficult. Indeed, one of the roles of the reviewer, being an expert in the domain, is to provide insight regarding the effects found or that could have been found. Often this insight is in the form of explanations for a predicted (or unpredicted) effect presumed to confirm the author's explanation, but that are instead different than the author's explanation. If one or more of these alternative explanations seem highly likely, the information value of the test may be so undermined as to make the hypothesis worthy of suppression. If the alternatives are merely possible, the reviewer should advise the recipient to note them in the discussion. Finally, there might be cases where an explanation for findings not predicted by the authors seems clear to the reviewer, but apparently unfamiliar to the authors who are perhaps less familiar with the literature. This situation creates a dilemma for the reviewer concerned about HARKing. That is, the reviewer might have found a pearl for the author to shine by connecting a result to an explanation. This prediction followed from the theory or previous findings and thus *the hypothesis existed prior to* the results being known (i.e., it was retrieved;

https://doi.org/10.1017/iop.2020.8 Published online by Cambridge University Press

Rubin, 2017), but the researchers did not know this (or understand it) when designing the study and analyzing the results. Surely, it behooves the reviewer to point it out.

Of course, one can avoid the integrity horn of this dilemma by suggesting that the retrieved hypothesis be presented as post hoc. Such a presentation is generally not considered HARKing (Kerr, 1998). In many cases this is a way to go because the results found are more curious than obvious. However, the cost of avoiding the integrity horn may sometimes mean being hoisted on the constructive horn. HARKing is *advocated* because it saves readers from wasting their time reading ill-conceived and ill-informed introductions and from hurting the reputation of researchers who are just trying to navigate the information overload that is our scientific literature (Bem, 2000). Indeed, research hypotheses are predictions of effects that can derive from theories. Periodically, the derivation is flawed (i.e., the prediction does not actually follow from the theory) and sometimes this error is caught by reviewers because they know the theory well. Does it serve the science to make sure the author's misunderstandings are made public? Moreover, if the hypothesis is retrieved from the existing literature, does exposing the ignorance of the author serve science? I do not see how.

A constructive purpose of the interaction between reviewers and authors is to find the highest information value of a study. I recall having recommended to authors that they restructure their article away from the original focus because the methods were too weak for drawing any reasonable conclusions regarding the original research question. Yet, some ancillary results seemed to challenge a different, newly emerging theory. Changing the focus required additional hypotheses and analysis to address alternative explanations for the challenging results. This advice was taken, and the article was published. Had the article retained the original introduction, including all the flaws the reviewers pointed out regarding the tests of the a priori hypotheses only to switch gears to talk about the challenged theory, it would have been much less efficient. Moreover, claiming the new hypotheses were post hoc would be incorrect. They existed before the study was conducted, it was just that the authors did not know that. Ironically, the possible ill of developing a new theory based on a spurious result would be less likely if reviewers helped researchers connect their results to existing theory and the predictions these theories make (cf., Kerr, 1998).

The second ill, suppressing a result and thus biasing the population of observed effects that can be sampled by meta-analysts, seems unavoidable, useful, and unnecessary. It is unavoidable for the reasons given above: poor research should be suppressed. Useful because poor quality research should not be included in meta-analysis because it biases the interpretations (e.g., a study with a serious confound in a *variable* will make the examined effect look different than what the effect of the *construct* being examined really is). It is unnecessary because suppressing a hypothesis does not mean the result is suppressed. It might still show up in the correlation matrix even if it is deemed unworthy of hypothesis status because of a change in article focus or some other reason (e.g., theoretical rationale too weakly reasoned; the prediction too unconnected to the theory being tested to justify retaining; prediction too robustly known to spend time defending in the introduction). What is clearly *not* good is suppressing a weak effect for an application (i.e., predictor or intervention study) when there is no clear methodological flaw. However, in cases where there are substantive flaws identified by reviewers, I assert that HARKing to suppress these findings does not disadvantage the aggregate knowledge available for meta-analysts.

The third ill, destined to repeat the past failures because of suppressed tests, is an ill the field should aspire to suffer. The issue is not likely to arise when a study is reasonably free of Type II threats and the finding of a null hypothesis reasonably disconfirms the theory. In that case, finding a null effect would and should be welcomed. However, concerns about explanations *other than an invalid theory*, at least as applied in the context of the study, might lead a reviewer to suggest suppressing the result. This *might* lead to inefficiencies if others repeat the same mistake, which *might* occur if the protocol would be tried by others ignorant of its shortcomings. Yet, the field's theories are so many, our methods so varied and weak, and the empirically challenging problems so numerous that few are likely to even think of tackling some research question the same way.

https://doi.org/10.1017/iop.2020.8 Published online by Cambridge University Press

Someday it might be useful to have a repository of failed protocols to check prior to sinking resources into a project, but that day does not seem imminent. For now, we can provide advice of a more generic nature (i.e., do not use cross-sectional, observational designs to test questions of causality) and seek to prevent turning our journals into a repository of ill-conceived ideas examined with weak protocols. Indeed, it seems much more likely that we have robust effects spuriously attached to explanations because the methods used are too weak to make valid conclusions. That is, we are probably not suppressing as many hypotheses as perhaps we should.

As applied psychologists, an issue of integrity is to not promote applications prior to a thorough scrutinizing of the intended and unintended consequences such an application might produce. At this point, I am concerned that the consequences of advising reviewers to refrain from encouraging "questionable research practices" might undermine, rather than enhance, the quality of our literature. To be sure, suggesting one report a p-value of .053 as $p < .05$ is simply not understanding the meaning of "less than" and should not be encouraged by reviewers. However, when it comes to HARKing, Köhler et al.'s (2020) competencies of expertise, representativeness, and constructiveness seem more relevant than this one element of the integrity competency.

Conclusion

Scientists are tasked with producing and disseminating knowledge. Efforts to get that knowledge out in a clear, understandable manner, likely with the help of reviewers and editors, seems to be part of the job. A cynical view is that scientists have displaced these noble pursuits with the pursuit of publications regardless of the validity of the knowledge presented in them. Perhaps an overly active propensity to trust leads me to think that most of us are trying to do the right thing by writing articles and helping others to write articles that reveal the truth as best as one can, given the current state of the literature, a dataset, and the design used to collect that data. I hope reviewers will continue to be allowed to support that effort via a well-reasoned set of guidelines and understandings, given the complexity of science.

References

Bem, D. J. (2000). Writing an empirical article. In R. J. Sternberg (Ed.), *Guide to publishing in psychology journals* (pp. 3–16). Cambridge, UK: Cambridge University Press.

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

Köhler, T., González-Morales, M. G., Banks, G. C., O'Boyle, E. H., 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.

Lipton, P. (2005). Testing hypotheses: Prediction and prejudice. *Science*, **307**, 219–221. doi: 10.1126/science.1103024

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, **21**, 308–320. doi: 10.1037/gpr0000128

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimentation designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Cite this article: Vancouver, J.B. (2020). Navigating the review process through the holier than thou. *Industrial and Organizational Psychology* **13**, 72–75. https://doi.org/10.1017/iop.2020.8

https://doi.org/10.1017/iop.2020.8 Published online by Cambridge University Press