Journal of Management
Vol. 43 No. 1, January 2017 5–18
DOI: 10.1177/0149206316679487

Editorial Commentary

Harking, Sharking, and Tharking: Making the Case *for* Post Hoc Analysis of Scientific Data

John R. Hollenbeck
Michigan State University
Patrick M. Wright
University of South Carolina

In this editorial we discuss the problems associated with HARKing (Hypothesizing After Results Are Known) and draw a distinction between Sharking (Secretly HARKing in the Introduction section) and Tharking (Transparently HARKing in the Discussion section). Although there is never any justification for the process of Sharking, we argue that Tharking can promote the effectiveness and efficiency of both scientific inquiry and cumulative knowledge creation. We argue that the discussion sections of all empirical papers should include a subsection that reports post hoc exploratory data analysis. We explain how authors, reviewers, and editors can best leverage post hoc analyses in the spirit of scientific discovery in a way that does not bias parameter estimates and recognizes the lack of definitiveness associated with any single study or any single replication. We also discuss why the failure to Thark in high-stakes contexts where data is scarce and costly may also be unethical.

Keywords: *philosophy of science; micro topics; research design; research methods; statistical methods; macro topics*

Study #1

A graduate student, desperate to get a job, takes 30 of the shortest and most easily obtained survey measures and creates a pair of long questionnaires. The first questionnaire includes 15 of the indices and categorizes them as independent variables (IVs), and the second

Corresponding author: Patrick M. Wright, Department of Management, Darla Moore School of Business, University of South Carolina, 1705 College Street, Columbia, SC 29208, USA.

E-mail: WrightJOM@moore.sc.edu

questionnaire includes the remaining indices categorized as dependent variables (DVs). The student then recruits 2,000 Mechanical Turk workers for pennies per hour, and sends out the questionnaires separated in time by 2 months. The respondents fill out the surveys in minutes in order to maximize their effort-to-pay ratio. At the end of the 2 months, the 30 × 30 correlation matrix generated by this process is analyzed, and this near-random data produces 20 correlations that are significant at the .05 level of probability level, some of which cross the Time 1 – Time 2 divide. Half of the correlations make no sense whatsoever; however, one could weave a plausible post hoc narrative that integrates theories from several different literatures to explain these results among the remaining ten statistically significant correlations. Some of these findings were totally unanticipated (and thus novel), and some were even counterintuitive (thus challenging the current knowledge base), and thus, the causal model that linked the 10 together might be well-received in journals that emphasize those two criteria.

The student converts some of the IVs and DVs to mediators based on the results and the post hoc narrative, and then presents the results as an a priori causal model that is written up and published in a top journal. The study attracts a great deal of attention, because of its novelty and counterintuitive nature, and several other research teams try to replicate the findings. None are able to do so, and many of these subsequent findings wind up unpublished because they were essentially reporting null results. Eventually, however, enough direct replications and indirect replications (i.e., reports of the parameter that were not directly intended as replications or part of a formal hypothesis) seep through the literature to allow a meta-analytic examination. This meta-analytic follow-up, based upon a sample size 30 times larger than the original study, fails to support any of the inferences reported in the original article and provides the best true estimates (near zero) of all the parameters that were part of the original study. Researchers in the field eventually abandon the model, and the field moves on to other models. Many people speculate on whether the graduate student was incompetent, unethical, or just very lucky, but in the end, everyone moves on to better, more robust models.

Study #2

A team of experienced epidemiologists suspect, based upon past published findings and well-established theory in their literature, that a certain drug might cure a novel life-threatening disease caused by a new virus. They secure funding from the National Institutes of Health, quickly recruit 100 patients from across the United States, and launch an experimental trial where half of the subjects are given the drug and half serve as controls. After two years of study, the results reveal a correlation of .10 between the treatment and survival, which, with this sample size, is not statistically significant.

One of the researchers notes to others that she is disappointed because she knows one woman in the eastern region where this researcher worked who was cured by the treatment. Another researcher chimes in and notes that she knew a woman in her region in the south who was also cured. When the third researcher from the north reported the exact same observation, no one waits for the fourth researcher from the west to tell his story—they are already reanalyzing the data. The results indicate that when analyzed separately by gender, the effect size for men is .00 and the effect size for women is .20, which, with this sample size, is still not statistically significant.

Disappointed, but not deterred, a discussion that lasts for days ensues regarding all the many different physiological differences between women and men that might explain this

result. All of these speculations are based upon the researchers' implicit knowledge of existing theory and empirical evidence in this area and are truly deductive in origin and a priori in spirit. Some of these are explored empirically with no luck. Eventually, this discussion focuses on how the drug might interact with estrogen levels due to its chemical composition, and the research team deductively arrives at a hypothesis aimed at testing the moderating effects of this variable. Because estrogen in women peaks at specific ages, the team goes back and reanalyzes the data broken down by age. They find that among women who are the peak age for estrogen levels, the correlation between the treatment and being cured is .50, which, even with this reduced sample size, is statistically significant. The authors immediately write these results up for publication as a short note in order to get these findings into the literature as soon as possible. In the short note, they write up the results for the age-by-gender interaction in the Discussion Section of their manuscript, noting that these were the result of an exploratory analysis of the data that was conducted after the main effects for the drug were found to be nonsignificant. They also schedule speaking tours at conferences, universities, and other laboratories in order to disseminate their results.

Their talks and presentations follow the *formula for a good detective story* (which this research was), and audiences are enthralled. The short note publication follows the *formula for a good scientific study* (which this research was) and is widely read. Researchers from across the globe immediately try to replicate their effects; and although 20 studies in 10 different countries are planned, after 10 studies involving 1,000 research participants, a meta-analytic summary reveals a correlation of .40—slightly smaller than the original .50 estimate—but with this much larger sample size, highly statistically significant. Due to ethical implications of not immediately treating the women in this age group who have this disease, the remaining 10 studies are discontinued and woman across the world are treated. Eventually, this discovery saves thousands of lives.

Introduction

Many scientific fields have recently been experiencing serious doubts about the reliability and validity of the empirical knowledge base on which their disciplines rest. Almost all of the research of stem cell researcher Hwang Woo Suk was found to be fraudulent (Wade & Sang-Hoon, 2006). Following an Office of Research Integrity investigation, Harvard evolutionary biologist Marc Hauser was forced to admit the finding that some of his papers contained fabricated data (Wade, 2010). In the social and behavioral sciences, social psychologist Dietrich Stapel confessed to having fabricated some of the data he reported in published studies (Bhattacharjee, 2013), and "Ego Depletion" theory has been called into question based on the failure to replicate its basic results across over 2,000 subjects in 24 simultaneous studies conducted across the globe (Engber, 2016).

The fields of management and work psychology have not been immune to these issues, and similar problems have been identified within this specific realm of the social and behavioral sciences. For instance, *Leadership Quarterly* recently retracted a number of articles it had published over the previous 5 years. Without describing in great detail the reasons for the retractions, Atwater, Mumford, Schriesheim, and Yammarino (2014) described the conditions that may justify a retraction. While certainly plagiarism and a violation of ethics justify such an action, they focused more on the replicability of findings, either through authors providing data for others to replicate their analyses, or providing detailed enough description

of the methodology for reviewers to assess contribution and others to replicate studies. They write that

> *Nature* and *Science*, the two most broad-based and widely-cited journals across all fields, have comparatively high retraction rates. These high retraction rates may, in part, arise because authors seek publication in high-visibility journals by any means, fair or foul. (p. 1179)

Taken together, all of this has led to a crisis in confidence regarding the scientific process. In fact, Fanelli (2013) wrote,

> Against an epidemic of false, biased, and falsified findings, the scientific community's defences are weak. Only the most egregious cases of misconduct are discovered and punished. Subtler forms slip through the net, and there is no protection from publication bias. (p. 149)

Although there are many issues related to the general collective anxiety being experienced by many in the behavioral and social sciences, one of the most pressing concerns regards the degree to which the parameters that we hold to be true, the inferences based upon those parameters, and the prescriptions we make for practice based upon those inferences are actually all false due to Harking (or HARKing). Hypothesizing after results are known (HARKing; Kerr, 1998), or accommodational hypothesizing (Hitchcock & Sober, 2014), refers to the process of retroactively presenting an unanticipated finding as if it were an a priori prediction or of failing to report a "failed" hypothesis. In other words, it describes the practice of evaluating an existing data set with a set of a priori hypotheses, and then either dropping those that were not supported and/or adding as a priori hypotheses relationships discovered to have been significant in that same data set.

One of the key standards for evaluating the validity of some theoretical proposition is the degree to which it allows one to predict the future. Prophecy is the ultimate test of true knowledge, and many extremely robust findings from our literature have stood the test of time. These can be reliably employed in the field and in the classroom where the results to nonprofessionals look like magic. Anyone who has used a $20 auction to teach people the dangers of "commitment to a failing course of action" knows the true predictive power of this proposition. Even though not every teacher can get someone to pay over $2,000 for a $20 bill (see Murnighan, 2002), that proposition never fails to secure bids well over $20. Moreover, the only thing more powerful for students than watching the phenomenon unfold is that the teacher knew for sure it was going to happen and built a lesson plan around it.

However, our field is also littered with propositions that failed to pass the test of time, and many false leads have been pursued at great expense of time and talent to no useful end. The field would have been better off had these false leads never been pursued, and many detective stories recount how the real suspects escaped detection due to the pursuit of false leads. The situation is made even worse if during some long and pointless pursuit, these ideas may have been presented to businesses or to students who will never take that specific course again and, hence, are forever misinformed.

Many now believe (Bosco, Aguinis, Field, Pierce, & Dalton, 2015) and, in fact, have long believed (Kerr, 1998), that some of these false leads are directly attributable to Harking. Study #1, described above, is a perfect example of how Harking could be employed to mislead the field. The parameters published from that study were all attributable to ambient

confounds and sampling error, and if held up to a proper test that stringently divided the probability by the number of actual tests conducted, none of them would have passed an appropriate test of statistical significance. However, because the results were presented as if they were a priori, the test as employed was far too liberal. These parameters were reported as genuinely different from zero, with all the ensuing inferential problems, not the least of which is that none of the parameters would ever replicate. Moreover, given the difficulty of publishing null results, it may take quite a bit of time before this inevitable failure to replicate becomes evident.

No reasonable or responsible person would look at Study #1 and claim that this reflects good professional practice. Many would also claim that the practice violated the ethical guidelines of the relevant professional societies, and was even immoral, depending upon the sophistication of the student. However, the research team described in Study #2 was also hypothesizing after the results were known given the strict dictionary definition of those terms. That is, although they relied on the existing empirical knowledge base to arrive at Drug A as a possible cure, the words "gender," "age," and "estrogen" never appeared in their National Institutes of Health grant proposal.

In addition, although when presenting their results in professional arenas, they clearly told the story exactly as it happened (i.e., a detective story that played out over a long time period); when they published this work as a short note, they employed the standard format for the target journal. Due to page length constraints, they did not get into a detailed recounting of the "discussion that lasted for days" regarding all the many different physiological differences between men and women other than estrogen that were considered and the subset of those that were empirically tested as alternative hypotheses. The reporting of the study was short, to the point, and told quickly enough to save lives. Most people would have considered *not* publishing the study as unethical, and any delay that prevented the treatment of women at risk in the specific age group from receiving treatment might be considered immoral. The value in the study laid more in the way in which it triggered the creation of a body of research much larger than the original study, as opposed to a pure estimate of any specific parameter. Indeed, the .10 difference between the parameter estimated in the original study and meta-analytic estimate may have been a small price to pay for triggering the larger program of research that eventually honed in best estimate.

Strictly speaking, however, the authors of both Study #1 and Study #2 were hypothesizing after some set of results were known given the dictionary definition of those words. Yet the fact that publishing Study #1 would be generally considered unethical, and failing to publish Study #2 would generally be considered unethical, suggests that there may be value in making distinctions between Harking in different ways, different contexts, and different places. The purpose of this article is to explore such distinctions and describe how to change the *formula for a good scientific study* in the field of management, so that one form of Harking is discouraged, while the other is not only encouraged but should be required as part of any thorough scientific inquiry. That is, we are going to discriminate the process of Sharking (Secretly Hypothesizing After the Results Are Known) and the process of Tharking (Transparently Hypothesizing After the Results Are Known) and argue that Tharking should be part of every published empirical study for any authors who do not have perfect ability to omnisciently predict the future. (We note that this latter recommendation is the opinion of the authors and is not meant to reflect the policy of the *Journal of Management*.)

That is, just as the Discussion section of all manuscripts should include sections on theoretical implications, practical implications, limitations, and future research, we will argue that Discussion sections should also include a section entitled "Post Hoc Exploratory Analyses." This would both reduce the motivation to engage in the reprehensible process of Sharking and encourage the valuable practice of Tharking. We believe changing the standard formula for how to report research results in this manner would make for a more honest and reliable set of scientific findings, as well as more engaging, thought-provoking, and informative Discussion sections. In the following sections, we (a) elaborate on the difference between Sharking and Tharking and (b) describe why in high-stakes contexts, where data are costly to obtain, and where the knowledge base is poorly developed, Tharking, unlike Sharking, has real value.

Tharking Versus Sharking

All normal science is, in a sense, hypothesizing after some results are known. Hence, the term Harking is somewhat a regrettable label for the phenomenon it attempts to describe because it presumes *how and where* certain actions took place, but those presumptions are not captured by the dictionary definition of those terms. That is, all normal science is a deductive process where past results and theories are used to make a priori predictions about unknown events. Ensuring that one's predictions are *informed, consistent with, and plausible relative to extant findings* is so critical to the process that failure to adequately make the case that past results support future prediction is a clear reason for rejecting a paper prior to reading the Method section. And to be clear, many papers get rejected well before the reviewer gets to the Method section, let alone the Results sections. Thus, if the post hoc narrative surrounding the Harking that took place in Study #1 described earlier was so convoluted that the review team felt that (a) two of the predictions flew in the face of all we know, (b) two made no sense at all, and (c) two were internally inconsistent with each other, then no level of statistical significance is going to make anyone want to publish such a study. Statistical significance may be a necessary condition for publication, but as we all know, it is very far from sufficient. Failure to fully place one's study well within the extant literature is thus such an important requirement of science that *not* hypothesizing after thoroughly knowing the results from past studies is unacceptable.

We define Sharking as "publicly presenting in the Introduction section of an article hypotheses that emerged from post hoc analyses and treating them as if they were a priori." This is redefining "Harking" as used in the current scientific literature to make it more consistent with dictionary definitions of the terms being employed (i.e., more specifically highlighting where and how this particular form of hypothesizing after the results were known took place). Thus, Sharking emphasizes a specific deviation from the normal scientific process because rather than hypothesizing after knowing the results from *all extant studies* (which is standard practice), the researcher hypothesizes after knowing the results from *the data at hand*.

As we noted earlier, Sharking distorts the meaning of certain statistical tests. Even more importantly, however, it is very easy to retrospectively make sense out of sampling error (see Schmidt, 1996). Within the realm of philosophy of science, evolutionary theories of causation (Collingwood, 1940) argue that human beings have evolved to be natural detectors of relationships because there was survival value in detecting causes—even if they were delayed, probabilistic, and highly contingent on third variables. Given this natural human

predilection, external safeguards against faulty causal detection need to be put in place, and the a priori standard for predictions is one such safeguard in science. One does not have to be an expert in evolutionary epistemology to know that "the future is difficult to predict—especially in advance." Thus, while it is essential to reflect all past results *except those currently at hand* into one's predictions, injecting the results currently at hand into one's predictions is always dangerous, and indeed unethical, when it is done in secret.

However, the question becomes whether the high absolute value placed in reflecting other people's past results into one's current predictions reverts to absolute zero when the results happen to be those currently in hand if this is done *transparently*. We define Tharking as "clearly and transparently presenting new hypotheses that were derived from post hoc results in the Discussion section of an article." The emphasis here is on how (transparently) and where (in the Discussion section) these actions took place. It entails developing new, post hoc, hypotheses to be tested with the data at hand and presenting their tests as such. As a process, Tharking allows one to adjust statistical tests and is a free admission that the "prediction" that was made and tested with the same data should not be treated with the same respect that one might afford to a purely a priori prediction. Still, with all the cards on the table, and informed by adjusted statistical tests, as well as a healthy skepticism regarding humans' ability to retrospectively make sense of almost anything, the results are presented tentatively. Thus, while accepting the proposition that there is zero value in Sharking, Tharking may not have a value of absolute zero, especially in high-stake contexts where data are costly to obtain (in terms of time, money, and effort from human subjects, investigators, funding agencies, and university support), and the potential value in terms of controlling certain outcomes is very high.

We also need to distinguish Tharking from other forms of post hoc analyses. For example, Tharking is not simply data mining, such as what we described in Study #1 at the outset of this article. That is, while Tharking entails post hoc analyses, it is driven by informed reconsiderations that prompt interesting questions. Could it be that the relationship exists for one sex, but not the other? Could it be that socioeconomic status suppresses the relationship? Thus, it entails developing new hypotheses to test rather than being purely data driven. Finally, it is not, as one reviewer stated, "wild-a-- guessing." The questions asked clearly must be based in theory and past results, albeit potentially theories that were not used in the development of the initial hypotheses and whose relevance only became obvious when confronted with some unanticipated finding.

Why Tharking May Have Value

Complexity and the Limits of Deductive Capacity

Although one might hopefully believe that the power of human deduction is infinite, in reality, the complexity of many real-world problems may preclude the ability to totally deduce all potential relationships. As Nobel Laureate Herbert Simon stated, "The capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behavior in the real world" (Simon, 1957: 198, 202). If left just to our powers of deduction, we as researchers may not just miss a lot, we may miss almost everything. Simon was using this as an argument for the value of modeling (mathematical, computational, and agent-based) to the study of

human and group behavior in order to capture complexity in a way that outstrips what can be done within a single human mind.

However, since the modeler controls both the inputs and processes that go into a simulation, and then engages in multiple runs, this gets very, very close to Tharking. Still, few would argue that simulations do not have their rightful place in the study of human behavior as both a check on the internal consistency of a formal deductive theory, as well as a speculative device that might generate insights that may not be deducible without a formal simulation. Like simulations, exploratory post hoc analyses based on Tharking may alert researchers to complex patterns or findings that could not be formally deduced by nonomniscient researchers without some aid.

Serendipity and the Science of Discovery

Despite clear textbook descriptions of the scientific process that paint a picture where all new results are understood solely in connection with past results, the history of science is replete with examples of accidental findings that simply could never have been anticipated by the extant knowledge base. In many cases, important discoveries arose strictly out of chance or from errors that created the level of variation needed to make a quantum leap change in the ability to predict or control some important phenomenon. This history would include the discovery of penicillin, plastic, X-rays, safety glass, saccharine, and even the Big Bang.

None of these discoveries or products were the outcome of formal deductive reasoning, but instead, all were generated by pure serendipity. Some estimates suggest that 33% to 50% of scientific discoveries were unexpected by-products of research formally studying some unrelated phenomenon (Dunbar & Fuselsang, 2005). In fact, beyond being unable to help us uncover facts that were never recognized before, because the deductive process relies so heavily on the existing knowledge base, it may very well work *against* generating new scientific discoveries. For example, as Thomas Kuhn (1962: 52) noted in his seminal treatise on scientific revolutions, "Edison's electric light bulb was produced in the face of unanimous scientific opinion that arc light could not be sub-divided." Thus, rejecting any findings that were not a product of a formal deductive process may limit our ability to detect new discoveries when the extant consensus in the literature is that something is impossible to anticipate.

Interestingly, the requirement of many journals to require "new discoveries" actually plays a part in encouraging Sharking. Authors that follow theory to develop logical hypotheses and test them may hear reviewers or editors suggest that the results were "too predictable." However, unexpected results require that they are … unexpected. Thus, authors are encouraged to present these unexpected results as if they were hypothesized in order to satisfy the deductive paradigm.

Historically Low Power in Primary Studies: Lessons From Meta-Analyses

Despite a long history of documentation that most studies in the social and behavioral sciences are underpowered due to small sample sizes, these fields have been amazingly steadfast in terms of refusing to increase the sample sizes (Cohen, 1992; McClelland, 1997). One might think that the major liability associated with underpowered studies, that is, failing to reject the null hypothesis (and a trip to the file drawer) might motivate researchers to avoid low power. However, as Maxwell (2004) showed, most studies are not nearly as

underpowered as one might think. Most admonitions regarding low power are based upon analyses that presume the researcher is testing one single a priori parameter.

However, most researchers are testing many, many hypotheses in a single study. Some of these hypotheses may reflect Sharking; however, this need not be the case. The presence of multiple a priori hypotheses, unadjusted for the number of tests, means that one can generate many statistically significant findings because the experiment-wise power is much higher than the power for any one specific a priori parameter (Maxwell, 2004). Clearly, if the number of hypotheses exceeds the number of research participants, then the Type I error rate is far too high (Hollenbeck, DeRue, & Mannor, 2006) for any one parameter. However, in many cases, the inflation of Type I error rates caused by this practice may be offset by the much higher rate of Type II errors that are avoided by limiting researchers to a single a priori hypothesis or an adjusted alpha for multiple hypotheses.

Meta-analytic evidence does suggest that when a relationship is framed as an a priori hypothesis, the correlation is roughly .06 points higher relative to when the exact same relationship is just reported as an indirect by-product of some study where that relationship was not the a priori hypothesis (i.e., indirect replications; see Bosco et al., 2015). One might accept this .06 estimate as the documented inflation of parameters created by Sharking and Tharking. However, because the sample size associated with meta-analyses are often 20 or 30 times larger than any one single study, the small loss in effect size over time due to Sharking or Tharking bias may be more than offset by the power of the statistical inference for rejecting the null when it is indeed false.

Maxwell, Lau, and Howard (2015), for example, showed how 30 attempted replications—all of which were not able to reproduce a previously reported statistical finding when evaluated by the criterion of whether or not the parameter was statistically significant in any of the single replications—were nonetheless totally supportive of the inference when subject to a meta-analysis where the sample size was 30 times larger than what was associated with any of the individual replication studies. Indeed, the required sample size for testing "nonreplicability" is much higher than people generally believe, and mere lack of statistical significance for a replication study is hardly definitive for drawing a different inference relative to the original study (Maxwell et al., 2015). As Maxwell et al. note,

> Designing appropriate replication studies frequently requires larger sample sizes than most researchers are accustomed to, or else the results of any single replication study are likely to be equivocal. … As a result, just as it may be unwise to consider a single original study as definitive, it may also be unwise to regard a single replication study as providing the final word. Instead, researchers should expect that multiple replication studies will often be needed to resolve apparent inconsistencies in the literature. (p. 495)

Research requirements that limit scientists to one single a priori prediction or adjust multiple predictions for the number of tests—post hoc or a priori—could have the perverse effect of creating null results and Type II errors that might have been avoided had a literature been given a chance to develop to the point where a meta-analysis could take place. Indeed, if we have learned any lesson from decades of meta-analyses, it is that most relationships, when assessed over long time periods, are much more statistically significant relative to what was suggested in the original underpowered studies (Schmidt, 1996). Tharking may create the opportunity to increase the true statistical power of our research and lower the rate of Type II

errors while, at the same time, preventing a large number of Type I errors that result from Sharking.

Grounded Theory and Qualitative Insights

Although not steeped in rich quantitative data, most approaches to grounded theory rely on the richness of qualitative observations in order to detect patterns associated with some phenomenon. This may be due to the stage of the research or the nature of the research, or simply the interests of the researcher; but again, few would deny the value for this kind of research to aid the process of discovery, despite the fact that it is not formally driven by top-down theoretical deduction. For example, Karl Weick's seminal research on sense-making was based on an admitted preoccupation with a single, $n = 1$ event (the Mann Gulch disaster) that was studied with an intensity that may be unmatched in the annals of social science (Weick, 1995).

The insights derived from this analysis could never be "replicated" in a strict technical sense. Indeed, with respect to future prediction, Weick (1995) himself noted that "I'm weak on boundary conditions, strong on shameless generalizing. Much of my work is basically an existence proof. If an event can happen in one place then it likely can happen again." Much in the same way, a researcher who was in a context with rich quantitative data and armed with an open mind might be able to use a rich set of empirical data to triangulate on a discovery that would not be deductively arrived at, but indeed "does happen again."

Clarity in Writing and Exposition

Sharking inevitably produces manuscripts where the Introduction sections are filled with tortured text that does not make any sense to the well-informed reader. Someone with a deep appreciation for the extant literature often understands implicitly that the current knowledge base is not adequate for explaining findings that result from Sharking efforts. Thus, when such a reader confronts an author that is Sharking, the reader's reaction to the Introduction is confused and incredulous. The reader may or may not infer that the work reflects Sharking, but in any event, the experience of reading such a manuscript is jarring and hard to integrate into one's own understanding of the knowledge base.

Donald Hambrick (2007), when suggesting that the field's devotion to top-down theorizing may have gone too far, noted this exact problem. When discussing the need to frame every discovery as if it were deducible from the extant knowledge base, Hambrick noted that

> the straightforward beauty of the original research idea will be largely lost. In its place will be what we too often see in our journal: a contorted, misshapen, inelegant product, in which an inherently interesting phenomenon has been subjected to an ill-fitting theoretical framework. (p. 1349)

In contrast, a manuscript that went into the research question with a framework that was consistent with the extant theory and literature, but then, in the Discussion section, Tharked on why what *should have happened in this context did not happen* in this context, would generate a much more authentic narrative. Readers, armed with the knowledge that the results were not predicted a priori, could evaluate the narrative on its own speculative

merits; and if deemed plausible, this may quickly generate a new study that can make a valid a priori prediction. Again, quoting Kuhn (1962: 49), "the prelude to much discovery and to all novel theory is not ignorance, but the recognition that something has gone wrong with existing knowledge." Tharking allows us to admit the existing knowledge base was found wanting and not pretend that it is flawless and always able to explain what will happen across diverse contexts.

The Ethical Costs of Failing to Leverage Costly Data

As we noted at the outset, Sharking is an unethical practice because of its outright intention to deceive the reader. Tharking, in contrast, is not an act of deception; and in many cases, when the problem is critical, and the data are expensive to obtain in terms of time, money, and effort, failure to Thark might also be considered unethical. That is, professional ethical guidelines reflect a pluralistic stance towards moral philosophy where the costs and benefits of various activities have to be weighed. For example, the cost of deceiving research participants of the true hypothesis being tested may be outweighed by the gains this practice may have for eliminating threats to valid inference, and hence, deception itself is not unethical per se, according to such standards.

In the same vein, if there is potential value in doing some specific post hoc exploratory analysis and it has almost no cost (e.g., rerunning the analyses separately for men versus women), then failing to do the analysis is unethical given the costs already paid by the university, the funding agency, the research assistants, the research participants, and the review team—especially since the practice does not rely on deception. Clearly, post hoc results have to be presented as such for all the reasons we have already noted. However, if the a priori predictions tested at great cost yield nothing of value, and if these could give way to post hoc predictions that may create value, then it is beyond lazy not to perform and report such analyses—it is unethical. Indeed, when it comes to "snooping" around the data, Rosenthal (1994) has directly stated that failing to snoop

> makes for bad science because while snooping does affect p-values, it is likely to turn up something new, interesting, and important. It makes for bad ethics because data are expensive in terms of time, money and other resources and because the anti-snooping dogma is wasteful of time, effort and other resources. If the research was worth doing, the data are worth a thorough analysis. (p. 130)

Action Steps

Although there is no justification for Sharking, we believe that there are justifiable reasons for Tharking. Indeed, an appreciation of the implicit advantages of Tharking can be seen in the behavior of review teams who often force authors to go back and revisit hypotheses and reanalyze data as a condition for revising and resubmitting an article. This is not an uncommon practice, and although this is technically "journal-mandated" Sharking (and hence, forcing authors to engage in unethical behavior), it does not *feel* like Sharking because it is so obviously public, transparent, and the review team *really does not necessarily know* a priori what the results from some post hoc analyses might be. This behavior is really Tharking by the review team and the authors working together for a perceived common good. Still, it

shares the problems with Sharking because most of this is secret to the eventual reader who would be better informed if this was treated as Tharking in the Discussion section, as opposed to going back and reframing the Introduction section.

An appreciation for Tharking, and exploratory research in general, can also be seen in the creation of the journal *Academy of Management Discoveries*, which was developed to serve as an outlet for pretheoretical and exploratory research that was not deemed acceptable in more traditional outlets. The mission of this journal is to "promote exploratory empirical research of management and organizational phenomena that our theories do not adequately explain," and the journal explicitly "welcomes studies at the pre-theory stage of knowledge development, where it is premature to specify hypotheses." Tharking takes this spirit and applies it to research that did not necessarily start out exploratory, but wound up going in that direction when that is what the evidence demanded.

In order to promote Tharking as a more ethical and efficient way of conducting research, we suggest the following courses of action for editors, action editors, reviewers, and authors. First, we believe that editors can discourage Sharking and promote Tharking by stating a clear policy that they prefer, or even require a section on "Post Hoc Analyses" where authors are able to discuss findings that they did not originally expect but that may be of interest in moving a particular field forward. The policy should encourage authors to develop their theory-driven hypotheses, test them, and report those results regardless of their significance. Particularly in cases where no significant findings emerge, authors should then be encouraged to transparently report their explorations for alternative findings that may spark future research. This section would also be the logical place to present Tharking that was done by the review team, and the editor should ensure that no author should ever be forced to engage in Sharking based upon a threat of rejection for failing to do so by an associate editor or reviewer.

In addition, editors can accept or even encourage high-quality quantitative inductive research. While the deductive paradigm has served our science well, it need not be the exclusive paradigm for conducting and publishing research. Editors will need to make clear policy statements of their openness to inductive quantitative research, and repeatedly emphasize to reviewers and action editors that such research, if conducted rigorously, falls within the domain of their journal.

Action editors (AEs) may hold the key to ensuring that journals promote Tharking over Sharking. AEs see the original submission as well as the comments from all reviewers. We have all experienced AEs who have suggested (usually based on one or more reviewers suggestions or demands) that the authors conduct additional analyses and are encouraged to include these as part of the hypotheses. Occasionally the authors do so without the direction of the AE, but nonetheless feel compelled to do so due to the nature of the AE's comments. Action editors should ensure that all Tharking be presented as such in the final version of the paper. The addition of a post hoc analyses section may make this easier to do. However, transparency can also be achieved by footnotes. Regardless of the means, it should be the primary responsibility of AEs to ensure that all Harking is actually Tharking.

Reviewers also can strongly positively or negatively impact whether or not authors Shark or Thark. As we previously mentioned, astute reviewers can often spot the twisted theoretical logic or lack of connection to existing literature in the development of hypotheses that, lo and behold, end up being supported. In such cases, reviewers must call out authors through either rejecting the papers or suggesting a massive revision that develops hypotheses clearly tied to

theory and literature, reporting those analyses, and then presenting their unique findings as part of a post hoc analysis. We hope that as explicit Tharking increases in perceived legitimacy, authors will be less prone to Sharking and reviewers will face fewer such manuscripts.

Second, reviewers asking for additional analyses should be clear with authors regarding how they want those reported. Again, with the addition of post hoc analyses sections, it will be easier for reviewers to direct authors to conduct analyses and transparently report them. When revised manuscripts present additional analyses as having been developed a priori, reviewers can and should reject papers for Sharking.

Authors should feel freed from the bondage of constraints imposed by a pure ratio-deductive paradigm to write more along the line of "good detective stories." All authors should begin with their preferred theoretical framework and develop the hypotheses that logically follow from within that framework. This should be both easier and more enjoyable than having to fit empirical square pegs into theoretical round holes. Then, having the freedom to explore, they can transparently report findings from their data that may be unique, interesting, or even startling, proposing them to future researchers for replication, extension, and more definitive meta-analytic parameter estimation.

Conclusion

We commend authors, editors, and others who have raised the warning flags regarding the potential dysfunctional outcomes of Harking. We agree that Sharking is both unethical and detrimental to the progress of science. However, we also believe that Tharking is beneficial to scientific progress and, in many cases, ethically required. We hope that all stakeholders to management research, editors, action editors, reviewers, and authors will heed our call to promote Tharking so that our field can become known for research that is effective, efficient, cumulative, and ethical.

References

Atwater, L. E., Mumford, M. D., Schriesheim, C. A., & Yammarino, F. J. 2014. Retraction of leadership articles: Causes and prevention. *Leadership Quarterly*, 25: 1174-1180.

Bhattacharjee, Y. 2013, April 26. The mind of a con man. *The New York Times*.

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. 2015. Harking's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology*, 2015: 1-42. doi:10.1111/peps.12111

Cohen, J. 1992. A power primer. *Psychological Bulletin*, 112: 155-159.

Collingwood, R. G. 1940. *An essay on metaphysics*. Oxford, UK: Clarendon Press.

Dunbar, K. N., & Fuselsang, J. 2005. Causal thinking in science: How scientists and students interpret the unexpected. In M. E. Gorman, R. D. Tweney, D. C. Gooding, & A. P. Kincannon (Eds), *Scientific and technical thinking*: 57-80. Mahwah, NJ: Lawrence Erlbaum Associates.

Engber, D. 2016. *Everything is crumbling*. *Slate*. Available from http://www.slate.com/articles/health_and_science/cover_story/2016/03/ego_depletion_an_influential_theory_in_psychology_may_have_just_been_debunked.html

Fanelli, D. 2013. Redefine misconduct as distorted reporting. *Nature News*, 494: 149. doi:10.1038/494149a

Hambrick, D. C. 2007. The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal*, 50: 1346-1352.

Hitchcock, C., & Sober, E. 2004. Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, 55: 1-34.

Hollenbeck, J. R., DeRue, D. S., & Mannor, M. 2006. Statistical power and parameter stability when subjects are few and tests are many: Revisiting the relationships between CEO personality and firm performance. *Journal of Applied Psychology*, 91: 1-5.

Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. *Personality & Social Psychology Review*, 2: 196-217.

Kuhn, T. S. 1962. *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Maxwell, S. E. 2004. The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9: 147-163.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, 70: 487-498.

McClelland, G. H. 1997. Optimal design in psychological research. *Psychological Methods*, 2: 3.

Murnighan, J. K. 2002. A very extreme case of the dollar auction. *Journal of Management Education*, 26: 56-69.

Rosenthal, R. 1994. Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science*, 5: 127-134.

Schmidt, F. L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1: 115-129.

Simon, H. A. 1957. *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*. New York, NY: John Wiley and Sons.

Wade, Nicholas. 2010, August 20. Harvard finds Marc Hauser guilty of scientific misconduct. *The New York Times*.

Wade, N., & Sang-Hoon, C. 2006, January 10. Researcher faced evidence of human cloning, Koreans report. *The New York Times*.

Weick, K. E. 1995. *Sensemaking in organizations*, Vol. 3. Thousand Oaks, CA: SAGE Publications.