2022

Questionable Research Practices Among Researchers in the Most Research-Productive Management Programs

Sven Kepes

Sheila K. Keener
Old Dominion University, skeener@odu.edu

Michael A. McDaniel

Nathan S. Hartman

WILEY   Journal of Organizational Behavior

Questionable research practices among researchers in the most research-productive management programs

Sven Kepes[1] | Sheila K. Keener[2] | Michael A. McDaniel[3] | Nathan S. Hartman[4]

[1]Department of Management and Entrepreneurship, Virginia Commonwealth University, Richmond, Virginia, USA

[2]Department of Management, Old Dominion University, Norfolk, Virginia, USA

[3]Department of Management and Entrepreneurship (emeritus faculty), Virginia Commonwealth University, Richmond, Virginia, USA

[4]Department of Management and Quantitative Methods, Illinois State University, Normal, Illinois, USA

Correspondence
Sven Kepes, Department of Management and Entrepreneurship, Virginia Commonwealth University, Richmond, VA, USA.
Email: skepes@vcu.edu

Summary

Questionable research practices (QRPs) among researchers have been a source of concern in many fields of study. QRPs are often used to enhance the probability of achieving statistical significance which affects the likelihood of a paper being published. Using a sample of researchers from 10 top research-productive management programs, we compared hypotheses tested in dissertations to those tested in journal articles derived from those dissertations to draw inferences concerning the extent of engagement in QRPs. Results indicated that QRPs related to changes in sample size and covariates were associated with unsupported dissertation hypotheses becoming supported in journal articles. Researchers also tended to exclude unsupported dissertation hypotheses from journal articles. Likewise, results suggested that many article hypotheses may have been created after the results were known (i.e., HARKed). Articles from prestigious journals contained a higher percentage of potentially HARKed hypotheses than those from less well-regarded journals. Finally, articles published in prestigious journals were associated with more QRP usage than less prestigious journals. QRPs increase in the percentage of supported hypotheses and result in effect sizes that likely overestimate population parameters. As such, results reported in articles published in our most prestigious journals may be less credible than previously believed.

KEYWORDS
Chrysalis Effect, HARKing, questionable research practices, research integrity

1 | INTRODUCTION

Publications in journals, especially top-tier journals, are a primary determinant of academic rankings (Ball, 2005; Nosek et al., 2012; Ostriker et al., 2009). In many disciplines, including management, a scholar's prestige is often assessed by considering the number of articles published in our most prestigious journals (Podsakoff et al., 2008). Thus, journal articles, especially in prestigious journals, are a key determinant of hiring, salary, pay raises, and grants, as well as promotion and tenure decisions (Gomez-Mejia & Balkin, 1992;

Nosek et al., 2012). Some manuscripts are more publishable than others, regardless of their scientific quality, and evidence suggests that management journals prefer articles addressing interesting and newsworthy topics with statistically significant results, preferably accompanied by new theory (Kepes & McDaniel, 2013).

"Publish or perish" is the phrase used to summarize the pressures faced by academics to publish. In response to this pressure, some researchers may see engagement in questionable research practices (QRPs) as instrumental to the goal of getting an article published (Kepes & McDaniel, 2013; Nosek et al., 2012). QRPs include activities

such as adding or removing hypotheses post hoc, adding or removing data to achieve statistically significant results, and selectively adding or removing variables (O'Boyle et al., 2017). Unfortunately, due to the almost institutionalized refusal to share data (Wicherts et al., 2006) and the difficulty publishing replication studies in our journals (Makel et al., 2012), there are both pressures to engage in QRPs and ample opportunities to do so.

Engagement in QRPs tends to be problematic as it can produce false positive results, which are costly scientific errors (Simmons et al., 2011). Essentially, QRPs can "convert Type I errors into non-replicable theory and hide null results from future generations of researchers" (Rupp, 2011, p. 486). To make matters worse, because our journals tend to shun replication studies (Makel et al., 2012), these false positive results persist and can distort our cumulative knowledge. Therefore, estimating the prevalence of QRPs is an important research endeavor. One approach to do this is to ask researchers about their own or others' QRPs (Banks, O'Boyle, et al., 2016; Banks, Rogelberg, et al., 2016; Bedeian et al., 2010). However, self-reported prevalence is likely a marked underestimate (John et al., 2012). A second approach to estimate the prevalence of QRPs is to compare an early description of a research project (e.g., a dissertation) with a later description of that project (e.g., a published article) (see O'Boyle et al., 2017; Pigott et al., 2013).

In this paper, we follow this second approach by building on the work of O'Boyle et al. (2017). Specifically, like O'Boyle and colleagues, we track research projects from the dissertation to journal publications and examine instances of omission (i.e., information present in a dissertation but not in an associated journal article) and commission (i.e., information not present in a dissertation but in the dissertation-derived journal article), which provides an estimate of early career QRP engagement. However, our study differs from O'Boyle et al.'s in several important ways. First, using a different, but complementary, theoretical lens, we develop a model grounded in well-established motivational perspectives (e.g., expectancy theory; Lawler, 1971) as well as testable hypotheses. Second, we assess the generalizability of O'Boyle et al.'s findings by examining two potential contingencies. Specifically, we explore the moderating effects of department research productivity and journal prestige. Third, we include potential QRPs that were not part of O'Boyle et al.'s study. Fourth, we study a well-defined population in contrast to O'Boyle et al.'s sample. Fifth, we conduct several sensitivity analyses on our results which allow us to have more confidence in our findings. Together, these differences allow us to assess the generalizability and potential boundary conditions of O'Boyle et al.'s study. These extensions make our study a constructive replication. Constructive replications are more than just literal replications, which rely on the exact same procedures (e.g., sample and research design) as prior studies (Köhler & Cortina, 2019). Rather, constructive replications extend and improve upon prior studies in an attempt to confirm (or disconfirm) the originally obtained results as well as answer an additional set of questions (Köhler & Cortina, 2019). This makes constructive replications a rare yet scientifically invaluable type of study (Grote & Cortina, 2018). As such, in addition to shedding new light on an important topic, our study answers numerous calls for replication studies in general (e.g., Banks et al., 2019; Banks, Rogelberg, et al., 2016; Simmons et al., 2011) and for constructive replications in particular (Grote & Cortina, 2018).

2 | HYPOTHESES DEVELOPMENT

As a theoretical underpinning of their study, O'Boyle et al. (2017) summarized general strain theory (Agnew, 1992), which views undesirable behavior, QRPs in this case, as stemming from negative social relationships or events, such as a failure to publish. They also propose that engagement in QRPs is more likely when one has the means, motive, and opportunity to do so. Building on these general ideas, we draw on expectancy theory (Lawler, 1971; Vroom, 1964) and propose that the reward system in academia motivates researcher behaviors. We rely primarily on expectancy theory for three reasons. First, the theory has a decades-long record of empirical support as well as the familiarity of researchers within the field of management. Second, expectancy theory has been previously applied to the context of our study (the reward system in academia; e.g., Estes & Polnick, 2012). Third, the expectancy theory framework subsumes theoretical arguments included in O'Boyle et al. (2017), including important aspects of general strain theory. For instance, O'Boyle et al. (2017, p. 378) noted that "germane to the Chrysalis Effect[1] is that the strain (experienced by faculty due to the blocking of desirable outcomes, such as the acceptance of a journal article) creates discomfort (similar to that of cognitive dissonance) that a person might attempt to remedy by use of nonideal channels of goal achievement (Kemper, 1978)." Cognitive dissonance is also an integral part of equity theory which has been integrated into expectancy theory (Lawler, 1971).

The environmental factors related to reward systems operate at both the department and journal levels. Specifically, the first category of environmental factors is the department-level reward systems that emphasize the quantity of publications, preferably in prestigious journals. We note that these departmental reward systems are enforced by departmental research committees as well as tenure and promotion committees. The second category of environmental factors involves journals' partiality for statistically significant results, preferences for theory development and for "hot" new topics (Campbell & Wilmot, 2018; Kepes & McDaniel, 2013), lack of data transparency requirements (Wicherts et al., 2006, 2011), and shunning of replications (Makel et al., 2012). We note that the second category of environmental factors tends to be enforced through a journal's editorial board and the reviewers. Thus, during the initial screening of submitted manuscripts and the editorial review process, journal editors and reviewers tend to ensure adherence to the second category of environmental factors. Our model thus posits that QRPs are motivated by department policies (e.g., financial and reputational rewards) and journal policies (see Figure 1). In addition, we propose that the extent to which environmental factors motivate researchers to engage in QRPs is moderated by the level of research productivity in the department in which a researcher resides as well as the prestige of the journal.
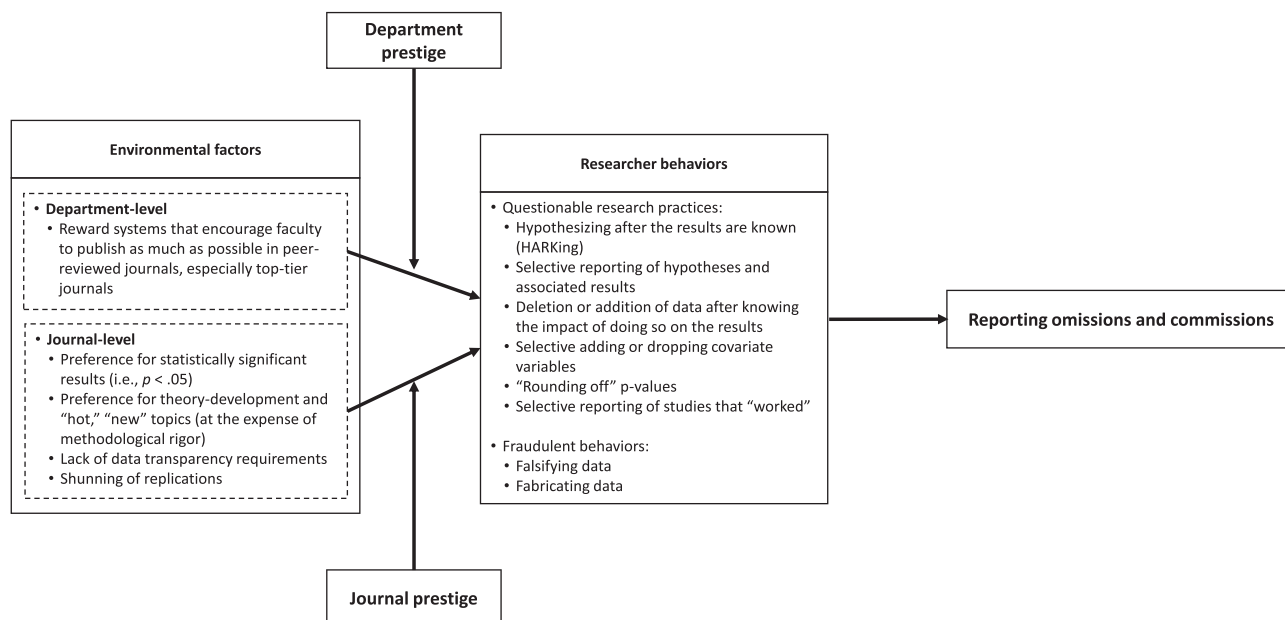
FIGURE 1 Environment factors, researcher behaviors, and reporting omissions and commissions

Expectancy theory suggests that individuals' motivation to engage in particular behaviors depends on their perceptions that their effort will lead to an anticipated level of performance (expectancy) and that this level of performance will lead to an outcome (instrumentality) that they value (valence). Because publications in our journals, especially the top-tier ones, are a primary determinant of academic rankings, departments emphasize the importance of faculty obtaining publications in those journals (Ball, 2005; Editors, 2006; Gomez-Mejia & Balkin, 1992; Nosek et al., 2012; Ostriker et al., 2009). Indeed, publications in these journals are a key determinant of several valued outcomes such as hiring, salary, and tenure decisions and may come with additional financial and reputational rewards (Editors, 2006; Gomez-Mejia & Balkin, 1992; Nosek et al., 2012; Podsakoff et al., 2008). Therefore, faculty know that publishing is necessary to obtain desired rewards (i.e., they have strong instrumentality perceptions).

Given journals' propensity for publishing articles with statistically significant results (Sterling & Rosenbaum, 1995), faculty are aware that their research efforts are more likely to yield a publication if their findings are statistically significant. They know they have the means to turn statistically insignificant results into significant ones, using behaviors such as adding or removing data and adding or removing covariates. They also know that they can remove unsupported hypotheses and add supported ones. Furthermore, they know that there is ample opportunity to engage in such behaviors and the possibility of detection (e.g., during the editorial review process) is rather small, meaning that such behaviors are unlikely to result in negative consequences (Banks, O'Boyle, et al., 2016; Banks, Rogelberg, et al., 2016; Bedeian et al., 2010). Consequently, faculty are likely to feel that if they engage in such behaviors, which are QRPs, they will be more likely to have their manuscript accepted for publication (i.e., they have strong expectancy [effort to performance] perceptions).

Taken together, faculty know that engaging in QRPs enhances the probability of successfully publishing a manuscript (expectancy), that publications are necessary to receive tenure and other rewards (instrumentality), and they value tenure and those other rewards (valence). They are thus motivated to display a wide range of behaviors and practices, including QRPs, to attain the statistically significant results necessary to publish in our journals in order to obtain the "prize of high pay" (Bloom, 1999, p. 28) and other rewards, such as tenure, promotions, grants, or a better position at some other university (Gomez-Mejia & Balkin, 1992; Kepes & McDaniel, 2013; Nosek et al., 2012). In addition, editors and reviewers occasionally request that authors engage in undesirable behaviors, such as QRPs (e.g., changing or dropping unsupported hypotheses), during the editorial review process (Rupp, 2011). Therefore, to have their publications accepted, and receive the rewards that accompany such publications, researchers who would otherwise not be inclined to engage in QRPs could be motivated to do so. Thus, the reward system in academia may be rewarding A (the use of QRPs) while hoping for B (the use of scientifically sound and rigorous processes and procedures) (Kerr, 1975).

Due to the reward system in academia and the resulting motivational pressures at the department and journal levels, authors may "polish" their dissertations into publishable articles in a number of ways. For instance, a published article may not contain information on a relation that was originally examined in the dissertation (in our model, a reporting omission). Conversely, an article may include a hypothesis about a relation that was not originally reported in the dissertation. This would be a reporting commission, especially if a post hoc hypothesis was presented as a priori, in which case it would constitute the QRP of hypothesizing after the results are known (HARKing; Kerr, 1998). Murphy and Aguinis (2019) identified one method of commission, which they

refer to as *question trolling*, defined as searching "through data involving several different constructs, measures of those constructs, interventions, or relationships to find seemingly notable results worth writing about" (p. 2). In a simulation of HARKing, Murphy and Aguinis found that this practice creates a large upward bias on the population effect size estimates when the pool of effects is large. However, it is possible that at least a few seemingly post hoc hypotheses in our dataset may not be based on trolling through the data as described by Murphy and Aguinis.[2] Therefore, we will refer to these hypotheses as *potentially HARKed*. In the context of turning dissertations into publishable journal articles, we thus predict:

> **Hypothesis 1.** Researchers are likely to disproportionately drop statistically nonsignificant dissertation hypotheses from the journal manuscript (H1a) and add statistically significant hypotheses to the journal article not found in the dissertation (H1b).

Another behavior that researchers can engage in to increase their chances of obtaining statistically significant results and, therefore, a publication, is to adjust the size of their sample. For instance, a researcher can stop a data collection effort as soon as a statistically significant result is obtained (Kepes & McDaniel, 2013). Similarly, researchers may be motivated to add data after they discover that a particular effect size is not yet statistically significant. Relatedly, a researcher can delete particular data points to move a "marginally" statistically significant result (e.g., $p < .1$) below the desired .05 threshold. Thus, researchers may alter the sample size associated with a result to obtain a statistically significant finding. More formally, we propose:

> **Hypothesis 2.** Depending on the level of statistical significance of a statistical test, researchers are likely to add (H2a) or drop (H2b) data before a journal article is published.

Researchers can also obtain statistically significant results through the use of covariates. Covariates can substantially influence obtained results—not only the choice of what covariate to include but also what operationalization to select (Becker, 2005; Carlson & Wu, 2011). For instance, firm size, a commonly used covariate in organization-level research, can be operationalized by the number of employees, the amount of total assets, or total sales, to name a few (Tosi et al., 2000). The choice of covariate constructs and their operational definitions can have a noticeable effect on whether an observed effect size reaches the desired level of statistical significance. Interestingly, about one fifth of researchers have indicated that adding or removing covariates was an appropriate practice (Banks, O'Boyle, et al., 2016). Thus, we propose that some researchers are likely to modify their covariates to achieve the level of statistical significance required for publication. More formally, we propose the following:

> **Hypothesis 3.** Depending on the level of statistical significance of a statistical test, researchers are likely to add (H3a) or drop (H3b) covariates before a journal article is published.

In addition, we predict that journal prestige moderates the frequency of presenting post hoc hypotheses as a priori hypotheses (HARKing; Kerr, 1998). For example, evidence from the medical sciences indicates that top-tier journals are more likely to contain statistically significant effect sizes than other journals (Easterbrook et al., 1991; Murtaugh, 2002). This is not particularly surprising. Publishing in highly prestigious journals is generally more competitive than publishing in less prestigious journals (Haensly et al., 2008), meaning that articles must stand out as particularly noteworthy to make it past the journal gatekeepers. As nonsignificant results are generally considered less interesting (Franco et al., 2014), articles with such results are less likely to stand out from the competition. Consequently, researchers should be more motivated to engage in QRPs to increase the percentage of supported hypotheses in their paper and thus the odds that their article is published in a more prestigious journal. Therefore, we predict the following:

> **Hypothesis 4.** QRPs such as HARKing (H4a), changing sample sizes (H4b), and changing covariates (H4c) are more common in more prestigious journals than in less prestigious journals.[3]

Similar dynamics may exist for departments in that the pressure to publish, especially in our most prestigious journals, is likely stronger in the most research-productive departments. As publications are a primary determinant of rankings (Ball, 2005; Nosek et al., 2012; Ostriker et al., 2009), to retain a given ranking, a school and department must ensure that its faculty continue to publish with great frequency. Otherwise, a school or department will slip in the rankings, which can have negative repercussions for, among others, fundraising activities and attracting desired new faculty and doctoral students. These dynamics also explain why tenure and promotion requirements are often significantly higher at the most research-productive departments when compared with less productive ones. These higher tenure and promotion requirements, in turn, suggest that faculty at highly research-productive departments may be more likely to engage in QRPs than individuals at less research-productive ones. Specifically, these faculty members already enjoy significant financial and reputational benefits compared with their colleagues at less research-productive departments. However, if tenure is not achieved, these researchers will likely lose access to these benefits and be forced to move to a less research-productive department. Prospect theory (Kahneman & Tversky, 1979) suggests that the framing of rewards as avoiding losses can significantly increase their motivating potential. Furthermore, evidence suggests that individuals are highly motivated to prevent the loss of status and reputation (Petit et al., 2010). Therefore, QRPs may be more likely in articles published by researchers from our most research-productive departments.

However, it is also possible that researchers at less research-productive departments may engage in more QRPs than researchers at the most research-productive departments. For instance, researchers at less research-productive departments could be especially motivated to display behaviors to increase their publishing record in order to obtain a position at a more research-productive department. Alternatively, individuals at more research-productive departments are likely to be well-trained and have access to more resources than researchers at less research-productive departments (e.g., access to companies for data collection and more funds to pay participants). Thus, they may be more prolific researchers who create better and more interesting research questions and design more rigorous studies. Therefore, they may not feel as motivated to engage in as many QRPs to get enough publications for tenure and other rewards. Due to these potentially competing dynamics and data constraints, we do not formally hypothesize this relation but, instead, pose the following research question:

> **Research question 1.** Do researchers at more research-productive departments display more (or less) QRPs than researchers at less research-productive departments?

In the course of coding data, we saw a distinction between essay and nonessay dissertations. Essay dissertations were defined as those that contained chapters that could stand alone as journal articles. When coding the essay dissertations and their associated journal articles, we found the results sections to be very similar and the wording in the dissertation chapter and the associated journal article was often identical in sections of both documents. Thus, we were curious if, as opposed to the comparison between nonessay dissertations and their associated journal articles, there would be fewer differences between essay dissertations and their associated journal articles. If there are substantial differences in the similarity between essay and nonessay dissertations and their resulting articles, dissertation type would be important to account for if one is attempting to estimate the prevalence of QRPs. This led us to offer the following research question:

> **Research question 2.** Do hypotheses from essay dissertations show less (or more) evidence of QRP engagement than hypotheses from nonessay dissertations?

Taken together, this study seeks to extend O'Boyle et al.'s (2017) research, which examined the extent to which hypotheses changed from dissertations to published articles, by focusing on researchers in the most research-productive management programs. Specifically, this paper examines the extent to which researchers in the top 10 management departments, as ranked by research productivity, engaged in QRPs by comparing the researchers' dissertations with journal articles based on those dissertations. We chose this sample of researchers because they tend to be highly productive, and their studies can be found in management's most widely read and highest impact scientific journals. They are also found on editorial boards and, as such, influence the direction of our field.

3 | METHOD

3.1 | Eligible dissertations

Table 1 shows the dissertation eligibility criteria that were determined a priori. We began with faculty members in the top 10 management programs,[4] as defined by the 2009–2013 rankings of management department research productivity developed and maintained jointly by the University of Florida and the Texas A&M University management departments. For this 5-year period, the list summed the number of publications in eight management journals (hereafter, the "Top 8 journals"). These journals were the *Academy of Management Journal*, *Academy of Management Review*, *Administrative Science Quarterly*, *Journal of Applied Psychology*, *Organizational Behavior and Human Decision Processes*, *Organization Science*, *Personnel Psychology*, and *Strategic Management Journal*. The top 10 research-productive management programs in this list were at the University of Michigan, University of Pennsylvania, Michigan State University, University of Maryland-College Park, Arizona State University, Harvard University, Pennsylvania State University, University of Minnesota, Twin Cities, Texas A&M University, and University of Texas at Austin. In seven of the 10 universities, the developers of the list identified one department as relevant to management; for three universities, two departments were identified (University of Michigan: Management and Organizations and Strategy; Harvard: Organizational Behavior and Strategy;

TABLE 1 Steps to determine eligible dissertations

Steps	k remaining
Step 1. Identify all faculty in top 10 research-productive management programs.	434
Step 2. Exclude teaching faculty and dean-level administrators.	303
Step 3. Date of dissertation is between 1994 and 2010, inclusive.[a]	117
Step 4. Dissertation was obtained.	113
Step 5. Dissertation had at least one explicit hypothesis that was evaluated empirically using a statistical test resulting in a *p* value.[b]	86
Step 6. At least one journal article, derived from the dissertation, contained a hypothesis that was evaluated empirically.	63

Note: k is the number of dissertations.
[a]We began our search with 1994 because that is when most dissertations from US universities became available online through ProQuest. We ended our search in 2010 to give time for the later dissertations to become journal publications.
[b]A "hypothesis that was evaluated empirically" was defined as a hypothesis whose support was determined through a statistical test that yielded a *p* value.

University of Minnesota: Work and Organizations and Strategic Management and Organizational Behavior). The screening continued through five additional steps (see Table 1).

3.2 | Nature of data

This study gathered data from archival records. Dissertations were obtained from ProQuest's *Dissertations & Theses Database* and interlibrary loan, when necessary. Journal articles judged to be drawn from dissertations were discovered through, and obtained from, online databases. An article was declared a match to a dissertation if the data set used in the article was substantially the same as in the dissertation, and the topic of the article was essentially the same as the topic of the dissertation.

3.3 | Data and unit of analysis

Our decision rules yielded 63 dissertations with one or more matching article(s). We note that none of the matching articles from the *Academy of Management Review* could be coded because no published article contained data or hypotheses. For dissertation-level analyses, there were 63 observations. At the hypothesis level, there were 2689 observations. A hypothesis-level observation could contain solely a dissertation hypothesis (1719 observations), a dissertation hypothesis and a matching article hypothesis (351 observations), or an article hypothesis alone (619 observations). Thus, we had a total of 2070 dissertation hypotheses and 970 article hypotheses. Our sample size for dissertation and article hypotheses is comparable with O'Boyle et al.'s article (1978 dissertation hypotheses and 978 article hypotheses).

Some dissertation or article hypotheses could be described as *combination hypotheses* in that they were a combination of discrete hypotheses, each with its own statistical significance test. Consider an illustrative example: *The Big 5 personality traits are positively correlated with job satisfaction*. This hypothesis could be evaluated with five statistical significance tests (one test for each Big 5 trait with job satisfaction). Because we sought to evaluate whether a hypothesis was supported based on a statistical test with $p < .05$, we broke down any "combination hypothesis" into its discrete individual hypotheses. Thus, the illustrative example would result in five observations in our data set. Although infrequent, in a given study (i.e., dissertation or journal article), a construct in a hypothesis was sometimes measured with multiple measures. For example, the Big 5 personality traits could be assessed by self-report measures and a peer rating. If the illustrative hypothesis were: *The Big 5 personality traits are positively correlated with popularity*, and there were two measures for each of the Big 5 personality traits and one measure for popularity, our data set would contain 10 observations.

Some hypothesis-level analyses were based on dissertation and article hypotheses that were the same (Tables 2a, 2b, and 4c). Table 3 analyses were based on all observations in the data set that met the criteria for the analysis. Finally, some analyses were based only on article hypotheses (Tables 4a and 4b).

3.4 | Sensitivity analyses

For our Table 2 series, we included two subsample comparisons: (1) only essay dissertations and (2) only nonessay dissertations. For Tables 4a and 4b, we analyzed all article hypotheses and provided additional analyses on three subsamples: (1) article hypotheses with at least one hypothesis matching a dissertation hypothesis, (2) essay dissertations, and (3) nonessay dissertations. Concerning Tables 4a and 4b, subsample 1, 14 of the 64 dissertation hypotheses had no hypotheses in common with their corresponding journal article(s). These 14 dissertations were clearly linked to their respective article(s) because the dissertation and the article used the same data set, and both sets of documents (i.e., dissertation and journal article[s]) were on the same topic. No hypotheses in common between a dissertation and an article derived from the dissertation occur when the article author(s) omitted all of the hypotheses originating in the dissertation and created new hypotheses based on the same data set and on the same topic for presentation in the journal article. For example, dissertation hypotheses could be moderator hypotheses (e.g., the positive relation between X and Y is moderated by M), but the matching article consisted of mediation hypotheses (e.g., the positive effect of X on Y is mediated by M). Some may be uncomfortable with our assertions that an article was derived from a dissertation but shared no hypotheses in common. Thus, this sensitivity subsample consists solely of dissertation and article hypotheses in which at least one hypothesis was common to both the dissertation and subsequent journal article.

The subsample analyses for Tables 2 and 4 serve as sensitivity analyses to assess the robustness of the obtained results from the full dataset. In addition to these sensitivity subsamples, we also include other sensitivity analyses.[5] Specifically, we examine whether our original results hold for other QRPs by examining the effect of scale manipulations (e.g., dropping items from a scale). We also use journal impact factor (JIF) as another indicator of journal prestige to examine the robustness of our originally obtained results. We then assessed whether (a) the number of articles derived from a dissertation and (b) the time between dissertation and article publication affected the motivational pressures on faculty and, thus, the prevalence of QRPs. Taken together, these sensitivity analyses, if supportive of original conclusions and theoretical predictions, allow one to have greater confidence in those conclusions. If conclusions differ, knowledge is gained by discovering boundary conditions of the conclusions.

3.5 | Risk ratios

Several hypothesis-level analyses use risk ratios, following the analysis approach of O'Boyle et al. (2017). A risk ratio is the ratio of ratios from two independent groups. In some tables, our two independent groups are (1) those dissertation hypotheses that are unsupported and (2) those dissertation hypotheses that are supported. In each group, we examine the hypotheses that changed in their support in the journal articles that were derived from the dissertations. For example, if there were 50 unsupported dissertation hypotheses but 5 of them became supported in the journal articles, the ratio for the

unsupported dissertation hypotheses group would be 5/50 or .10. If, in the supported dissertation hypotheses group, there were 100 supported dissertation hypotheses but 5 were unsupported in the journal articles, that ratio would be 5/100 or .05. To calculate the risk ratio, one would divide .10 (the ratio from the unsupported dissertation hypothesis group) by .05 (the ratio from the supported dissertation hypotheses group) resulting in a risk ratio of 2.00. This risk ratio of 2.00 shows that the unsupported dissertation hypotheses are twice as likely to become supported in the journal articles as are the supported dissertation hypotheses to become unsupported in the journal article.

3.6 | Data analysis considerations for hypothesis-level analyses

A risk ratio greater than 1.0 is consistent with an inference of the use of QRPs. Because all risk ratio's in O'Boyle et al.'s (2017) paper were above 1.0 (all were also above 2.0, actually), our analyses in Tables 2a and 2b used one-tailed statistical significance tests in anticipation of risk ratios well above 1.0. Because of the one-tailed tests, a risk ratio may be statistically significant even if the 95% confidence interval includes a value of 1.0 or less.

4 | RESULTS

Our results are first presented for the dissertation-level analyses and then for the hypothesis-level analyses.

4.1 | Dissertation-level analysis results

Hypothesis 1a suggested that researchers are likely to disproportionally drop nonsignificant dissertation hypotheses in subsequent journal articles. This analysis excludes seven dissertations in which all dissertation hypotheses were supported and thus is based on 56 (63–7 = 56) dissertations. In these 56 dissertations, there were 1310 unsupported dissertation hypotheses of which only 136 appeared in an article. That is, only 10.4% of the unsupported dissertation hypotheses appeared in a journal article. Of all 1310 unsupported dissertation hypotheses in the 56 dissertations, the median percent of nonsupported dissertation hypotheses reported in the articles was only 8.2% and the mean was 24.2%. To put this into context, we compare these findings to analogous findings for supported hypotheses. The supported hypothesis findings are based on 62 dissertations, as one dissertation only included unsupported hypotheses. In these 62 dissertations, there were 760 supported hypotheses, 215 (28.3%) of which appeared in a journal article. Thus, a greater percentage of supported dissertation hypotheses (28.3%) than unsupported dissertation hypotheses (10.4%) appeared in journal articles. Furthermore, of the 760 supported dissertation hypotheses in the 62 journal articles, the median percent of supported dissertation

hypotheses reported in the articles was 25.8% and the mean was 35.2% (compared with 8.2% and 24.2%, respectively, for unsupported dissertation hypotheses). Taken together, this shows that more non-supported dissertation hypotheses were excluded from the published article than supported dissertation hypotheses, which represents a reporting omission, providing support for Hypothesis 1a.

Hypothesis 1b suggested that another approach to increasing the percentage of statistically supported hypotheses in a journal article is to add new post hoc hypotheses for the article that are statistically supported and present them as a priori (i.e., HARKing). One way to estimate the prevalence of potentially HARKed hypotheses is to determine the percentage of dissertations that result in articles containing statistically supported hypotheses that do not appear in the associated dissertation. Only 6 of the 63 dissertations have yielded articles in which all of their hypotheses matched a dissertation hypothesis. This means that 57 (90.5%, 57/63) dissertations contain at least one potentially HARKed hypothesis in the corresponding journal articles. Of the 57 dissertations, four had zero supported article hypotheses with no matching dissertation hypothesis. Thus, this analysis was based on 53 (57–4 = 53) dissertations. For these 53 dissertations, the percent reported is based on the number of statistically significant article hypotheses that have no matching dissertation hypothesis divided by the number of journal article hypotheses in the article(s) associated with the dissertation. The percentages range from 7.1% to 100.0% (median = 41.7%, mean = 48.2%). Thus, a substantial percentage of supported journal article hypotheses has no corresponding dissertation hypothesis. This result supports Hypothesis 1b and is consistent with the inference that a substantial number of journal article hypotheses in these data are potentially HARKed.

4.2 | Hypothesis-level analysis results

To enhance the clarity of our results presentation, we have divided our discussion of the hypothesis-level analysis results into seven sections. The first section provides some summary statistics at the hypothesis level. Section 4.4 defines the overall Chrysalis Effect and compares the magnitude of this effect in our paper to the comparable value in O'Boyle et al.'s (2017) paper. Section 4.5 examines hypotheses common to a dissertation and a journal article to make inferences about the frequency of engagement in QRPs. The fourth section (Section 4.6) examines hypotheses from the dissertation that are not found in the derived articles as well as hypotheses from articles that are not found in the dissertation. This section permits inferences about the suppression[6] of results from dissertations and the likelihood of HARKed hypotheses in journal articles. Section 4.7 addresses the differences in QRP prevalence between Top 8 (i.e., more prestigious) journals versus non-Top 8 (i.e., less prestigious) journals in our sample as well differences related to department research productivity. The last two sections (Sections 4.8 and 4.9) detail results of two sensitivity analyses we performed on our data.

4.3 | Summary statistics

The data set had 2070 dissertation hypotheses based on 63 dissertations. Of these, 760 (36.7%) were supported (i.e., statistically significant at $p < .05$ in the hypothesized direction). Of the 970 article hypotheses, 577 (59.5%) were supported. There were 351 (36.2% of the overall 970) article hypotheses that matched a dissertation hypothesis, of which 216 (61.5%) were supported. A total of 619 (63.8% of the overall 970) article hypotheses did not match any dissertation hypothesis. Of these, 361 (58.3%) were supported. Because the 619 articles were published after the dissertation analyses were completed, they might reasonably be classified as post hoc, or HARKed, hypotheses. However, that is not how they are characterized in any of the articles. As noted earlier, we refer to these hypotheses as *potentially HARKed* to allow for the possibility that some were not HARKed. In summary, dissertation hypotheses are supported at a much lower rate (36.7%) than are article hypotheses (matching hypotheses: 61.5%; nonmatching hypotheses: 58.3%; all article hypotheses: 59.5%). Also, we note that the majority (63.8%) of the 970 article hypotheses in our data set might reasonably be classified as potentially HARKed.

4.4 | Overall Chrysalis Effect

O'Boyle et al. (2017, p. 377) labeled the use of QRPs to improve the probability of manuscript publication as the Chrysalis Effect with a metaphorical reference of "an ugly caterpillar (initial results) turns into a beautiful butterfly (journal article)." We begin with a presentation of the overall chrysalis results by comparing O'Boyle et al.'s results with ours. O'Boyle et al. defined an overall Chrysalis Effect by comparing the ratio of supported to unsupported hypotheses in the dissertations to the value of the same statistic in the articles. They concluded that the ratio more than doubled (2.4) in the journal articles (.82 to 1.94). In our data, the ratio is very similar (2.5; .58 to 1.47). In summary, one can infer that O'Boyle et al.'s (2017, p. 388) conclusion that "the published literature, at least as it relates to those early career efforts by junior faculty, is overstating its predictive accuracy by a substantial margin" is consistent with our results.

4.5 | Hypotheses common to both dissertation and journal article

Tables 2a and 2b are modeled after O'Boyle et al.'s (2017; see Table 1, p. 385) analyses. In these tables, we include the number of dissertations that contributed data to the analysis and present comparisons for those observations in which a dissertation hypothesis and an article hypothesis were the same. By the *same*, we mean that both the dissertation hypothesis and the article hypothesis made the same prediction. We refer to these pairs of hypotheses as common or matching hypotheses. Results displayed in Table 2a use all the data for which a dissertation hypothesis matches an article hypothesis.

Table 2b examines the data in two sensitivity subsamples (essay dissertations and nonessay dissertations). In the last column of Table 2a, we provide O'Boyle et al.'s risk ratio results for comparison purposes. In Table 2a, in the row labeled "all data," we see that there were 136 dissertation hypotheses that were unsupported (i.e., were not statistically significant at $p < .05$) and 215 dissertation hypotheses that were supported (i.e., were statistically significant at $p < .05$), yielding a total of 351 dissertation hypotheses that matched a journal hypothesis. Thus, 61.3% (i.e., 215/351) of the common hypotheses were supported in the dissertation

If QRPs influenced these data, there should be a greater probability of an unsupported dissertation hypothesis becoming a supported journal article hypothesis than the probability of a supported dissertation hypothesis becoming an unsupported article hypothesis. Of the 136 unsupported dissertation hypotheses, 22 (16.2%) were supported in the journal article derived from the dissertation. Of the 215 supported dissertation hypotheses, 21 (9.8%) were not supported in the journal article derived from the dissertation. Thus, 6.4% (16.2%–9.8%) more article hypotheses became supported for unsupported dissertation hypotheses than article hypotheses becoming unsupported for supported dissertation hypotheses.

These percentage differences can also be expressed as a risk ratio. Given that the risk for the unsupported dissertation hypotheses to become supported is .162 and the risk for the supported dissertation hypotheses to become unsupported is .098, the ratio of these two risks (i.e., the risk ratio) is 1.66 (.162/.098 = 1.66).[7] A risk ratio of 1.0 indicates no difference in ratios. Risk ratios above 1.0 are consistent with an inference of engagement in QRPs. This risk ratio is statistically significant ($p < .05$). Therefore, our results provide support for Hypotheses 1a and 1b. The comparable risk ratio in O'Boyle et al.'s (2017) paper is 4.52 and is significant at $p < .001$.

The row for add/drop data concerns changes in sample size. If the article hypothesis had a larger sample size than the dissertation hypothesis, data were added. If the article hypothesis had a smaller sample size than the dissertation hypotheses, data were dropped. The risk ratio was 3.20 ($p < .01$), supporting the inference that adding or dropping data was a QRP in this data set. This supports Hypothesis 2. We note that this finding was similar to the results for O'Boyle et al. (2.41, $p < .01$). In addition, we found a significant risk ratio for adding data (8.69, $p < .01$) but not for dropping data (2.13, $p > .05$), which supports Hypothesis 2a but not 2b. As such, the statistical support for add/drop data is primarily driven by the adding of data. Although O'Boyle et al.'s risk ratios were not statistically significant for adding data, they were statistically significant for dropping data. We note that all risk ratios for adding data and dropping data were above 2.0 and thus substantial in their magnitude, suggesting an effect supporting the QRP hypotheses, but some fell short of statistical significance.

Next, we examined the adding or dropping of covariates. If an article hypothesis used covariates not found in the dissertation hypothesis analysis, covariates were added. If the article hypothesis analysis did not use covariates that were used in the dissertation analysis, covariates were dropped. Unlike changes in sample size, in which

TABLE 2a QRPs statistics for hypotheses common to both dissertation and journal article for all hypotheses using risk ratios

QRPs and common hypotheses	Total N	ND	Unsupported dissertation hypotheses			Supported dissertation hypotheses			% diff.	RR (95% CI)	O'Boyle et al. (2017) RR (95% CI)
			N	$\Delta_{support}$	%	N	$\Delta_{support}$	%			
All data	351	49	136	22	16.2%	215	21	9.8%	6.4%	1.66 * (0.95, 2.89)	4.52*** (2.69, 7.60)
Add/drop data	216	37	90	16	17.8%	126	7	5.6%	12.2%	3.20** (1.37, 7.46)	2.41** (1.16, 5.01)
Add data	73	24	35	8	22.9%	38	1	2.6%	20.2%	8.69** (1.14, 65.97)	2.14 (0.59, 7.75)
Drop data	143	24	55	8	14.5%	88	6	6.8%	7.7%	2.13 (0.78, 5.82)	2.53* (1.04, 6.18)
Add/drop covariates	242	36	90	16	17.8%	152	20	13.2%	4.6%	1.35 (0.74, 2.47)	NA
Add covariates	242	36	90	16	17.8%	152	20	13.2%	4.6%	1.35 (0.74, 2.47)	NA
Drop covariates	108	25	42	9	21.4%	66	5	7.6%	13.9%	2.83* (1.02, 7.86)	NA
Change scale	83	15	33	8	24.2%	50	5	10.0%	14.2%	2.42* (0.87, 6.77)	NA

Note: In columns 11 and 12, the statistical significance value is for a one-tailed test consistent with expectation of a risk ratio greater than 1. For column 12, we recalculated O'Boyle et al.'s (2017) risk ratios in order to obtain the statistical significance value for a one-tailed test. $\Delta_{support}$ under column 5 is the number of statistically significant article hypotheses that are part of the common pair with an unsupported dissertation hypothesis. $\Delta_{support}$ in column 8 is the number of statistically insignificant article hypotheses that are part of the common (matching) pair with a supported dissertation hypothesis. The risk ratio is the likelihood of a nonstatistically significant dissertation hypothesis having a statistically significant article hypothesis compared with the likelihood of a statistically significant dissertation hypothesis having a nonstatistically significant article hypothesis. A risk ratio of 1 indicates no relationship. A risk ratio greater than 1 indicates support for the inference that article hypotheses statistical support was influenced by engaging in QRPs. Data are in the row analysis if the QRP was observed. The Add/Drop data row contains the observations of the hypotheses in which the article either adds data to sample or dropped data to the sample size. Note that the results for Add/drop covariates and Add covariates are the same. This is due to the finding that whenever covariates were not the same, a covariate was added (i.e., added covariates never had a zero value when covariates in the dissertation and article hypotheses were the not same). Table S5 contains summary statistics for continuous covariate variables.

Abbreviations: CI, confidence interval; N, the number of hypotheses in the group with group being unsupported dissertation hypotheses or supported dissertation hypotheses; ND, number of dissertations; QRPs, questionable research practices; RR, risk ratio; Total N, total number of hypotheses in the analysis.

*p < .05. **p < .01. ***p < .001.

one could add data *or* drop data, covariates could be dropped *and* added. Neither add/drop covariates nor add covariates had statistically significant risk ratios; however, dropped covariates had a risk ratio of 2.83 which was statistically significant (p < .05). Thus, Hypothesis 3b is supported although 3a is not. Comparisons to O'Boyle et al.'s results are not available because they did not examine covariate changes as potential QRPs. In summary, our data indicated that the QRPs of changing sample sizes (add/drop data and add data) and covariates (dropping covariates) occurred. Furthermore, the risk ratios that were not statistically significant were in the expected direction.

Table 2b presents results separately for essay and nonessay dissertations. Comparisons to O'Boyle et al.'s (2017) results are not possible because they did not compare essay to nonessay dissertations. There were only eight essay dissertations, and there were no differences in hypothesis support between the dissertation hypotheses and the article hypotheses for these dissertations. Given the number of zero cells in the analyses relevant to essay dissertations, we do not present risk ratios. For the analysis of the nonessay dissertations, we note that removing eight essay dissertations from this sensitivity subsample sharply reduces the number of supported dissertations hypotheses (163 for the nonessay dissertations compared with 215 in Table 2a). In Table 2b, the nonessay dissertation results mirrored the Table 2a results in that the risk ratios for

add/drop data, add data, and drop covariates were statistically significant. Similar to Table 2a, the statistically nonsignificant risk ratios were in the direction supporting the use of QRPs.

There is an important caveat concerning the data used for the Table 2 series. As noted earlier, in the all data row of Table 2a, 61.3% of the dissertation hypotheses in the matched sample are supported. However, as noted in summary statistics results, only 36.7% of the total 2070 dissertation hypotheses were supported. Thus, the matched dissertation hypotheses are clearly not representative of all the dissertation hypotheses in this study. We note that O'Boyle et al.'s data had the same issue. In their table, similar to our Table 2a, 57.8% of the common hypotheses were supported (O'Boyle et al., 2017, p. 384) yet only 44.9% of the 1978 total dissertation hypotheses were supported (O'Boyle et al., 2017, p. 388). We do not know how this difference affects results in the Table 2 series. However, the common hypotheses are the only way to test the effect of sample changes and covariate changes as QRP causes of hypothesis support changes from dissertation to article(s).

We also examined the potential effects of scale manipulations. Thus, we essentially replicated Hypotheses 2 and 3 with psychological scales and asked whether, depending on the level of statistical significance of a statistical test, researchers are motivated to alter at least one psychological scale involved in testing a hypothesis before the journal article is published. As shown in Tables 2a and 2b, this did

TABLE 2b QRPs statistics for hypotheses common to both dissertation and journal article for selected hypotheses

QRPs and common hypotheses	Total N	ND	Unsupported dissertation hypotheses			Supported dissertation hypotheses			% diff.	RR (95% CI)
			N	$\Delta_{support}$	%	N	$\Delta_{support}$	%		
Essay dissertations										
All data	63	8	11	0	0.0%	52	0	0.0%	0.0%	NA
Add/drop data	25	6	3	0	0.0%	22	0	0.0%	0.0%	NA
Add data	8	5	1	0	0.0%	7	0	0.0%	0.0%	NA
Drop data	17	5	2	0	0.0%	15	0	0.0%	0.0%	NA
Add/drop covariates	33	7	7	0	0.0%	26	0	0.0%	0.0%	NA
Add covariates	33	7	7	0	0.0%	26	0	0.0%	0.0%	NA
Drop covariates	13	5	1	0	0.0%	12	0	0.0%	0.0%	NA
Change scale	0	0	0	0	0.0%	0	0	0.0%	0.0%	NA
Nonessay dissertations										
All data	288	41	125	22	17.6%	163	21	12.9%	4.7%	1.37 (0.79, 2.37)
Add/drop data	191	31	87	16	18.4%	104	7	6.7%	11.7%	2.73** (1.18, 6.34)
Add data	65	19	34	8	23.5%	31	1	3.2%	20.3%	7.29** (0.97, 55.04)
Drop data	126	19	53	8	15.1%	73	6	8.2%	6.9%	1.84 (0.68, 4.98)
Add/drop covariates	209	29	83	16	19.3%	126	20	15.9%	3.4%	1.21 (0.67, 2.20)
Add covariates	209	29	83	16	19.3%	126	20	15.9%	3.4%	1.21 (0.67, 2.20)
Drop covariates	95	20	41	9	22.0%	54	5	9.3%	12.7%	2.37* (0.86, 6.54)
Change scale	83	15	33	8	24.2%	50	5	10.0%	14.2%	2.42* (0.87, 6.77)

Note: See Table 2a for notes related to the abbreviations used in the rows and columns. O'Boyle et al. (2017) did not compare essay to nonessay dissertations. Thus, they did not report corresponding RRs.
*$p < .05$. **$p < .01$.

appear to be the case, as the risk ratio was 2.42 ($p < .05$). Thus, our originally obtained results are robust to other QRPs, such as altering psychological scales to measure a construct of interest.

4.6 | Hypotheses not common to both dissertation and journal article

The results presented in Tables 2a through 2b are based on the portion of our sample where the dissertation hypotheses matched their corresponding journal article hypotheses. Table 3 presents the analysis of unique hypotheses that are either dissertation hypotheses with no matching journal article hypothesis (a dropped dissertation hypothesis) or journal article hypotheses with no matching dissertation hypothesis (an added article hypothesis). The latter journal article hypotheses ($N = 619$) are considered potentially HARKed hypotheses, because the dissertation data may have been reanalyzed to locate statistically significant results (question trolling; Murphy & Aguinis, 2019), which were then presented as a priori hypotheses.

Table 3 shows that 58.3% of these added article hypotheses were significant. Of the 1719 dropped dissertation hypotheses, only 31.7% were statistically significant. The risk ratio summarizing these relationships was 1.84 in our data and 1.81 in the corresponding

O'Boyle et al. (2017) data. These results support Hypothesis 1, both H1a and H1b, in that both risk ratios are statistically significant.

4.7 | The moderating effects of journal prestige and department research productivity

Tables 4a and 4b provide estimates of the probability of journal article hypotheses being potentially HARKed. We defined a hypothesis as potentially HARKed when the article containing the hypothesis was matched with a dissertation and the hypothesis is not included in the dissertation. We present the results separately for articles from the Top 8 journals, the set of prestigious journals used to rank the management departments by the University of Florida and the Texas A&M University management departments, and for articles not from the Top 8 journals. In addition to the results for the full data set, we also present the results for three sensitivity subsamples.

The left side of Table 4a presents the results for all 970 article hypotheses examined in this study. Of these 970 hypotheses, 619 (63.8%) meet our decision rule for potentially HARKed hypotheses. Of the 619 potentially HARKed hypotheses, 58.3% are statistically significant. If one looks just at the 624 hypotheses published in the Top 8 journals, the percentage of potentially HARKed hypotheses increases from 63.8% to 69.7% (435 of 624). Of the

TABLE 3 QRPs statistics among hypotheses appearing only in the dissertation or only in the journal article

QRPs and unique hypotheses	N	Significant	Percentage	% diff.	RR (95% CI)	O'Boyle et al. (2017) RR (95% CI)
Added journal article hypotheses	619	361	58.3%	26.6%	1.84*** (1.69, 2.01)	1.81*** (1.64, 1.91)
Dropped dissertation hypotheses	1719	545	31.7%			

Note: In column 6, the statistical significance value is for a one-tailed test consistent with expectation of a risk ratio greater than 1. To obtain the statistical significance for the risk ratio in column 7, we recalculated O'Boyle et al.'s (2017) risk ratio in order to obtain the statistical significance value for a one-tailed test.

Abbreviations: CI, confidence interval; QRPs, questionable research practices; RR, risk ratio.

***$p < .001$.

TABLE 4a Potentially HARKed journal article hypotheses

All article hypotheses		Article hypotheses with at least one hypothesis from an article matching a dissertation hypothesis	
Total article hypotheses	970	Total article hypotheses	787
Potentially HARKed journal hypotheses	619	Potentially HARKed journal hypotheses	436
Percent of total	63.8%	Percent of total	55.4%
Confidence interval of percent	60.7%–66.8%	Confidence interval of percent	51.8%–58.9%
Percent of potentially HARKed hypotheses that are statistically significant	58.3%	Percent of potentially HARKed hypotheses that are statistically significant	59.4%
Confidence interval of percent	54.3%–62.2%	Confidence interval of percent	54.6%–64.0%
All article hypotheses published in Top 8 journals		**Article hypotheses published in a Top 8 journal with at least one hypothesis from an article matching a dissertation hypothesis**	
Total article hypotheses	624	Total article hypotheses	462
Potentially HARKed journal hypotheses	435	Potentially HARKed journal hypotheses	273
Percent of total	69.7%	Percent of total	59.1%
Confidence interval of percent	65.9%–73.3%	Confidence interval of percent	54.4%–63.6%
Percent of potentially HARKed hypotheses that are statistically significant	60.7%	Percent of potentially HARKed hypotheses that are statistically significant	64.5%
Confidence interval of percent	55.9%–65.3%	Confidence interval of percent	58.4%–70.1%
All article hypotheses excluding those in Top 8 journals		**Article hypotheses not published in a Top 8 journal with at least one hypothesis from the article matching a dissertation hypothesis**	
Total article hypotheses	346	Total article hypotheses	325
Potentially HARKed journal hypotheses	184	Potentially HARKed journal hypotheses	163
Percent of total	53.2%	Percent of total	50.2%
Confidence interval of percent	47.8%–58.5%	Confidence interval of percent	44.6%–55.7%
Percent of potentially HARKed hypotheses that are statistically significant	52.7%	Percent of potentially HARKed hypotheses that are statistically significant	50.9%
Confidence interval of percent	45.3%–60.1%	Confidence interval of percent	43.0%–58.8%

435 potentially HARKed hypotheses in Top 8 journal articles, 60.7% are statistically significant. By contrast, if one looks at the hypotheses published in non-Top 8 journals, only 53.2% (184 of 346) are potentially HARKed, and of those, only 52.7% are statistically significant. One can compare the overlap in comparable confidence intervals to make judgments about the statistical significance of the difference in potentially HARKed hypotheses between Top 8 and non-Top 8 journals. For example, the confidence intervals for the percent of potentially HARKed hypotheses for Top 8 journals (65.9%–73.3%) do

not overlap with the comparable confidence interval (47.8%–58.5%) for non-Top 8 journals, and therefore, the corresponding percentages 69.7% and 53.2% are statistically significantly different ($p < .05$). Thus, there is a larger percentage of potentially HARKed hypotheses in articles published in the Top 8 journals than non-Top 8 journals, supporting Hypothesis 4a. The results for nonessay dissertations (see Table 4b) are largely consistent with the results for the full sample of journal articles. Specifically, the percent of potentially HARKed hypotheses was consistently higher for the Top 8 journals than for

TABLE 4b Potentially HARKed journal article hypotheses

Article hypotheses associated with essay dissertations		Article hypotheses associated with nonessay dissertations	
Total article hypotheses	121	Total article hypotheses	849
Potentially HARKed journal hypotheses	58	Potentially HARKed journal hypotheses	561
Percent of total	47.9%	Percent of total	66.1%
Confidence interval of percent	38.8%–57.2%	Confidence interval of percent	62.8%–69.2%
Percent of potentially HARKed hypotheses that are statistically significant	82.8%	Percent of potentially HARKed hypotheses that are statistically significant	55.8%
Confidence interval of percent	70.1%–91.0%	Confidence interval of percent	51.6%–59.9%
Article hypotheses published in Top 8 journals and associated with essay dissertations		**Article hypotheses published in Top 8 journals and associated with nonessay dissertations**	
Total article hypotheses	88	Total article hypotheses	536
Potentially HARKed journal hypotheses	49	Potentially HARKed journal hypotheses	386
Percent of total	55.7%	Percent of total	72.0%
Confidence interval of percent	44.7%–66.1%	Confidence interval of percent	68.0%–75.7%
Percent of potentially HARKed hypotheses that are statistically significant	81.6%	Percent of potentially HARKed hypotheses that are statistically significant	58.0%
Confidence interval of percent	67.5%–90.8%	Confidence interval of percent	52.9%–63.0%
Article hypotheses not published in a Top 8 journal but associated with essay dissertations		**Article hypotheses not published in a Top 8 journal but associated with nonessay dissertations**	
Total article hypotheses	33	Total article hypotheses	313
Potentially HARKed journal hypotheses	9	Potentially HARKed journal hypotheses	175
Percent of total	27.3%	Percent of total	55.9%
Confidence interval of percent	13.9%–45.8%	Confidence interval of percent	50.2%–61.5%
Percent of potentially HARKed hypotheses that are statistically significant	88.9%	Percent of potentially HARKed hypotheses that are statistically significant	50.9%
Confidence interval of percent	50.7%–99.4%	Confidence interval of percent	43.2%–58.4%

TABLE 4c QRPs usage comparison by journal prestige (Top 8 journal)

			Top 8 journals		Non-Top 8 journals		
QRP	Total *N*	*ND*	*N*	Percent (95% CI)	*N*	Percent (95% CI)	% difference
Add/drop	339	48	182	72.0% (64.8%–78.2%)	157	54.1% (46.0%–62.0%)	17.8%***
Add data	339	48	182	25.3% (19.3%–32.3%)	157	17.2% (11.8%–24.2%)	8.1%
Drop data	339	48	182	46.7% (39.3%–54.2%)	157	36.9% (29.5%–45.0%)	9.8%
Add/drop covariates	351	49	189	74.1% (67.1%–80.0%)	162	63.0% (55.0%–70.3%)	11.1%*
Add covariates	351	49	189	74.1% (67.1%–80.0%)	162	63.0% (55.0%–70.3%)	11.1%*
Drop covariates	351	49	189	49.7% (42.4%–57.1%)	162	8.6% (5.0%–14.3%)	41.1%***
Change scale	157	24	84	51.2% (40.1%–62.2%)	73	42.5% (31.2%–54.6%)	8.7%

Note: "Add/drop covariates" results match the "Add covariate" results because, when covariates did not match in common hypotheses analyses, at least one covariate was added.

Abbreviations: CI, confidence interval; *ND*, number of dissertations; QRPs, questionable research practices.

*$p < .05$. **$p < .01$. ***$p < .001$.

the non-Top 8 journals, although the difference was not always statistically significant.

Table 4c explored whether the QRPs of changing sample sizes and covariates were more prevalent in the top journals than in the journals not in the Top 8. As predicted in Hypotheses 4b and 4c, four

QRPs related to the changing of sample sizes and covariates were more common in the Top 8 journals than in the journals not in the Top 8. Statistical differences between Top 8 and non-Top 8 journals were found for add/drop data, add/drop covariates, add covariates, and drop covariates. The remaining two QRPs (add data and drop

data) were in the predicted direction but not statistically significant. These results provide some support for H4b and strong support for H4c.

We also conducted these analyses using the JIF as an indicator of journal prestige. For a description of how this analysis was conducted, please see the supporting information. These results were largely consistent with our initial results. Specifically, the percentage of potentially HARKed hypotheses that were statistically significant was larger in articles published in journals with JIF ≥ 3 at the time the article was published, though this difference was not always statistically significant. Furthermore, the QRPs of adding/dropping data, adding data, and dropping covariates were statistically significantly more likely to occur in journals with higher impact factors. Thus, our originally obtained results regarding journal prestige are generally robust to alternate conceptualizations of the measure of prestige. The complete results are presented in our supporting information (Tables S1a–c).

We also proposed that the research productivity of the department moderates the relation between department-level variables and QRPs. We explored this research question by comparing our results to O'Boyle et al.'s (2017). Our sample only contains researchers from the most research-active management departments, whereas O'Boyle et al.'s sample represents management departments with varying levels of research productivity, yielding an average of less research-active management departments when compared with our sample. Also, although O'Boyle et al. did not have data related to essay dissertations and the use of covariates, they presented results related to some of the other QRPs we examined. As discussed throughout our results section, our risk ratios were sometimes, but not always, larger in magnitude than O'Boyle et al.'s. Furthermore, the confidence intervals of our risk ratios and O'Boyle et al.'s overlap. Thus, we see no compelling support for a department's research productivity moderating the use of QRPs. However, the prestige of the journal is associated with the use of QRPs (Table 4c).

4.8 | Comparison of dissertations with one or more associated articles

As an additional sensitivity analysis, we re-examined some of our key results by comparing dissertations with one associated article to those with more than one associated article. We reasoned that the motivational dynamics related to QRP engagement may be stronger when only one article is published from a dissertation than when there are multiple articles published from a dissertation (e.g., the pressure and, thus, motivational force may be stronger if a dissertation yields only one published article). This did indeed appear to be the case. Specifically, a lower proportion of unsupported dissertation hypotheses appeared in subsequent journal articles when there was only one article derived from the dissertation (8.6%) compared with more than one article (18.7%). Importantly, the number of supported hypotheses included in subsequent articles is noticeably higher both when there is only one article derived from a dissertation (22.7%) and more than

one article derived from a dissertation (40.9%). We also examined the instances of other QRPs by the number of articles derived from a dissertation. With the exception of adding data and changing scale, the QRPs examined in our analysis were statistically significantly more likely to occur when only one article stemmed from the dissertation. The full results of these sensitivity analyses are available in the supporting information (Tables S2 and S3). This finding is in line with our initial theoretical supposition that motivational pressures lead to QRP engagement.

4.9 | Comparison of dissertations by number of years between dissertation and article publication

Finally, we explored whether the time between the dissertation and subsequent article publication affected the results. Given that most faculty apply for tenure during their fifth year, we reasoned that the motivational force of engaging in QRPs may be stronger for untenured faculty. Thus, we reran analyses regarding QRP usage comparing the results for articles published less than 5 years after the dissertation to those published 5 or more years after the dissertation. For more than half of the QRPs examined (4/7; 57.1%; i.e., add/drop data, drop data, drop covariates, and change scale), results indicated that the motivational pressure before year five is stronger than during or after year five. This provides additional evidence that the reward system in academia, particularly the prospect of obtaining tenure, motivates QRP engagement. The full results are in our supporting information (Table S4).

5 | DISCUSSION

Researchers face pressure to publish; their careers depend upon it. Because scientific journals tend to publish articles with statistically significant results and decline to publish articles containing statistically nonsignificant ones (Fanelli, 2012; Sterling & Rosenbaum, 1995), expectancy theory (Lawler, 1971; Vroom, 1964) suggests that researchers are motivated to use QRPs to increase their chances of successful publication. When engagement in QRPs, sometimes at the behest of editors and reviewers, yields a successful publication, the feedback researchers receive from rewards associated with the publication reinforces their use. As a result, the published scientific literature may be distorted and lacking in credibility, which negatively impacts science, teaching, and practice (Kepes et al., 2014b). In this study, we sought to extend the work of O'Boyle et al. (2017) by focusing on researchers in top management programs, as judged by research productivity. We chose to study researchers in the 10 most research-productive US management programs because they substantially influence our available literature. Their research productivity is typically very high, and as such, their research is likely overrepresented in journals relative to researchers in less research-active management programs. Furthermore, these researchers tend to serve as the gatekeepers of our journals due to their positions as editors and

reviewers. In these roles, they can request that authors engage in a variety of QRPs (e.g., adding data, changing scales, dropping or adding hypotheses) during the editorial review process. As such, examining their propensity to engage in QRPs is an important scientific endeavor.

Our findings supplement O'Boyle et al.'s (2017) findings in several ways. First, in contrast to O'Boyle et al., we analyzed changing covariates as a potential QRP and found evidence supporting our hypotheses. Relatedly, we also examined the effects of scale manipulations as a QRP and found results in line with our theorizing. Second, we found that essay dissertations appear to be quite different from other dissertations with respect to engagement in QRPs. There are several potential explanations of this finding. For instance, some chapters from essay dissertations may have already been under review by the time the dissertation was accepted and QRPs may have been introduced as part of the editorial review process. Alternatively, as an anonymous reviewer suggested, essay dissertations may be more focused in scope and thus contain a smaller number of carefully developed "core" hypotheses. More thoughtfully crafted dissertation hypotheses would likely mean that essay dissertations require less polishing when being prepared as submissions as journal articles than those submissions emanating from traditional dissertations. Third, we analyzed potentially HARKed journal hypotheses in articles by Top 8 journal status and showed that the articles in the Top 8 journals tended to have higher rates of potentially HARKed hypotheses. Fourth, we examined whether articles published in Top 8 journals have higher rates of other QRPs. We found that QRPs related to the changing of sample sizes and covariates were more common among Top 8 journals than in less prestigious journals. Taken together, these findings indicate that empirical results published in management's most prestigious journals could be less credible than findings published in less prestigious journals.

Our dissertation-level analysis results are consistent with three inferences. First, a large percentage of dissertation hypotheses are excluded from journal articles derived from the corresponding dissertations. These suppressed dissertation results (i.e., reporting omissions) are likely to receive less attention because they are not published. This causes publication bias in some of our literature areas, leading to misestimated meta-analytic means (Banks et al., 2015; Kepes et al., 2012). Second, the suppression of dissertation results in our sample is not a random suppression in that predominantly nonsupported dissertation hypotheses tend to be suppressed from journal articles. This likely causes an overestimation of effect sizes in the published literature, the most prevalent form of publication bias (Kepes et al., 2014a). Third, in our sample, unsupported dissertation hypotheses are often excluded from subsequent journal articles and these journal articles also include hypotheses not found in the dissertation (*potentially* HARKed hypotheses). One may argue that dissertations almost always contain more hypotheses than published articles. Thus, some dissertation hypotheses may have to be dropped for corresponding articles to get published (e.g., due to length restrictions). However, the issue is not that hypotheses are being dropped between a dissertation and the corresponding journal article(s); the

issue is that substantially more unsupported hypotheses than supported hypotheses are being dropped. Thus, the dropping of hypotheses is not random; instead, it is systematic, which systematically suppresses small effect sizes and, therefore, biases the publicly available scientific evidence (Kepes et al., 2012).[8]

At the level of the hypothesis, results from the Table 2 series provide support for add/drop data, add data, adding and deleting covariates, and altering scales as QRPs predictive of changing hypothesis support. These QRPs were statistically significant in both Table 2a (all common hypotheses) and Table 2b (for all nonessay hypotheses).

With regard to the role of journal prestige in our model, we obtained results generally consistent with our predictions. We found that our most prestigious journals contain a significantly larger percentage of potentially HARKed hypotheses than our less prestigious ones. We documented that articles published in our top journals are associated with more QRP engagement related to the changing of sample sizes and covariates than other journals. Therefore, empirical results in our most prestigious journals may be less credible than results presented in other journals. This should be concerning as it is counter to the commonly held belief that our most prestigious journals contain our very best empirical evidence. Comparing our results to O'Boyle et al.'s, we found no evidence that scholars from the top research management programs (our data) are more likely to engage in QRPs than average researchers in less research-productive departments (O'Boyle et al.'s data). Specifically, although our risk ratios for add/drop data and add data (see Table 2a), as well as add/drop hypotheses (see Table 3), are larger than those reported in O'Boyle et al., the confidence intervals overlap. This suggests that all management researchers are similarly motivated to engage in QRPs due to environmental factors.

One could speculate on the value of this study given that the seminal paper on this approach to analyzing QRPs was published by O'Boyle et al. in 2017. However, our study is a constructive replication, which extends and improves upon O'Boyle et al.'s study. Specifically, in addition to replicating most of O'Boyle et al.'s findings, we offer three additional contributions. First, relative to O'Boyle et al., we have a more well-defined population (i.e., researchers in the top 10 most research-productive management programs). Our sample includes journal editors (or past editors) and members of editorial boards. Second, we used a different conceptual model than O'Boyle et al. O'Boyle's paper relied on general strain theory as their theoretical framework. This theory was developed and is primarily used in criminal justice research. We used an expectancy theory framework, one of the most well understood and supported theories in management, to develop testable hypotheses. Third, in contrast to O'Boyle et al., we analyzed changing covariates and scale manipulations as potential QRPs, examined a potentially important methodological moderator (i.e., type of dissertation), analyzed the frequency of QRPs in more/less prestigious journals.

Overall, our findings were largely consistent with our hypotheses. Our hypotheses related to particular QRPs authors may engage in when moving from dissertations to articles (e.g., Hypotheses 2 and 3)

were largely supported. Although we could not test authors' motivations behind any changes that were made between dissertations and their published articles (e.g., authors may use these QRPs in response to requests during the editorial review process), the changes that we did observe suggest that scholars in our sample appear to have made a variety of choices aimed at increasing the likelihood that their papers would be published.

We also proposed that department-level and journal-level factors motivate individuals to engage in QRPs. Indeed, the reward system in academia incentivizes and, for tenure purposes, necessitates publications, especially top-tier ones. Furthermore, journals tend to publish predominantly statistically significant results (Sterling & Rosenbaum, 1995) and few replications (Kepes & McDaniel, 2013). These department-level and journal-level factors influence authors' expectancy, instrumentality, and valence perceptions associated with the use of QRPs. Stated differently, because departments emphasize publications and journals emphasize new and statistically significant

findings, researchers are part of a system which motivates them to engage in behaviors that will increase their number of statistically significant results and, thus, their chances of publication. Therefore, it appears that, although departments and journals are likely hoping to publish scientifically sound research that makes contributions to science and practice, they may actually be rewarding the use of QRPs. Table 5 provides a summary of our results related to the QRPs and contingency effects we examined.

Finally, we note that we conducted an array of sensitivity analyses to assess whether our originally obtained results are robust. For instance, we examined the effects of dissertation format (essay vs. nonessay dissertation), a methodological moderator, as well as several moderators related to motivational pressures on researchers (e.g., the effect of time between the dissertation and published article on the obtained results). Furthermore, we used alternative variables and measures to see whether our results were due to a particular variable or operationalization (e.g., changing of a

TABLE 5 Summary of QRPs examined and their research support

QRP	Research finding(s)
Drop unsupported dissertation hypotheses from the corresponding article	Supported by this article and by O'Boyle et al. (2017) for the change in hypothesis support (i.e., changed from an unsupported dissertation hypothesis to a supported article hypothesis).
Add a hypothesis to the corresponding article that was not in the dissertation	Supported by this article and by O'Boyle et al. for the change in hypothesis support (i.e., changed from an unsupported dissertation hypothesis to a supported article hypothesis).
Add or drop data (i.e., increase or decrease the sample size)	Supported by this article and by O'Boyle et al. for the change in hypothesis support (i.e., changed from an unsupported dissertation hypothesis to a supported article hypothesis). Also, in this article, this QRP is predictive of an article being published in a Top 8 journal.
Add data (i.e., increase the sample size)	Supported by this article for the change in hypothesis support (i.e., changed from an unsupported dissertation hypothesis to a supported article hypothesis). Not statistically significant in the O'Boyle et al. article, but the results are in the hypothesized direction. Not significantly predictive of an article being published in a Top 8 journal but in the predicted direction.
Drop data (i.e., reduce the sample size)	Supported by O'Boyle et al. for the change in hypothesis support (i.e., changed from an unsupported dissertation hypothesis to a supported article hypothesis). Not statistically significant in this article, but the results are in the hypothesized direction. Not significantly predictive of an article being published in a Top 8 journal but in the predicted direction.
Add or drop covariates	Predictive of an article being found in a Top 8 journal. Not a statistically significant predictor in the change of a hypothesis support (i.e., changed from an unsupported dissertation hypothesis to a supported article hypothesis), but the results are in the predicted direction. Not addressed in O'Boyle et al.
Add covariates	Predictive of an article being found in a Top 8 journal. Not a statistically significant predictor in the change of a hypothesis support (i.e., changed from an unsupported dissertation hypothesis to a supported article hypothesis), but the results are in the predicted direction. Not addressed in O'Boyle et al.
Drop covariates	Predictive of a change in hypothesis support (i.e., changed from an unsupported dissertation hypothesis to a supported article hypothesis). Predictive of an article being found in a Top 8 journal. Not addressed in O'Boyle et al.
Change scale	Predictive of a change in hypothesis support (i.e., changed from an unsupported dissertation hypothesis to a supported article hypothesis). Not addressed in O'Boyle et al.

Abbreviation: QRPs, questionable research practices.

psychological scale as an alternative QRP, JIF as an alternative measure for journal prestige). In doing so, we uncovered important dynamics and nuances (e.g., the differences between essay and non-essay dissertations) and determined that our originally obtained results were robust to alternative conceptualizations and operationalizations of variables. Therefore, we are confident in the accuracy of our results and associated conclusions.

5.1 | Limitations and recommendations for future research and practice

Although the results of our study were robust to several sensitivity analyses and make several contributions to the literature, it did have some limitations. First, given that the articles are derived from dissertations, our data reflect early career behavior. We do not know whether the QRPs observed in these data remain constant across the researchers' careers or get less or more frequent over time. Although one of our sensitivity analyses found that QRPs were more common in articles published within 5 years of one's dissertation, we cannot be sure if this result would generalize to articles that are not derived from one's dissertation.

Second, given the time frame of our study (dissertations between 1994 and 2010), it is possible that our results reflect norms regarding the studied QRPs that existed at the time the dissertations and articles were published and that authors in a more recent sample would not have engaged in these activities, even if pressured to do so by department and journal policies related to the reward system. We note though that O'Boyle et al.'s (2017) paper included a slightly more recent sample (dissertations published between 2000 and 2012) and found similar results to ours. Therefore, we recommend that our analyses be conducted using a more recent sample of dissertation-article pairs. Given that one has to provide enough time for dissertations to become articles, realistically, a more recent sample that examines current norms (particularly a well-defined sample) may not be available/large enough for several years.

Third, our sample is US-centric because our sampling frame only included a sample of the most research-productive programs in the United States. It is possible that the same levels of QRPs that we observed in this US-centric sample do not generalize to non-US samples. Therefore, we recommend that future research tests our general model and hypotheses using samples outside of the United States. Fourth, our data did not allow us to examine whether the environmental pressure from the department or the journal is stronger. Future research could request the set of comments and response letters associated with every published article to examine the role of the review process in supporting QRPs in a more nuanced fashion.

In addition to those recommendations presented above that address study limitations, we also offer suggestions for future research more generally. First, researchers should reproduce the current results using the data and syntax (data, R syntax, and decision rules are available on the Open Science Framework [http://doi.org/10.17605/OSF.IO/QHMWB]).. Second, the effect of culture variables

on the decision to include or exclude unsupported dissertation hypotheses could be examined. It may be that some departmental cultures encourage or discourage the inclusion or exclusion of such hypotheses. Thus, one might contrast departments with a strong research culture with departments that have a weak one. A third related recommendation is to examine the prevalence of QRPs as a function of the departments or schools in which the researchers were trained. Some graduate programs may encourage the use of QRPs more than others. We suggest a cycle by which a doctoral student or an early career faculty member might learn about QRPs from modeling behaviors of successful peers, mentors, or colleagues. Thus, scholars may be socialized to see engagement in QRPs as necessary for successful publication, which inevitability becomes part of their recommendations to other authors when they serve as reviewers and journal editors (i.e., a vicious cycle).

Fourth, comparisons of QRP prevalence in different research areas might indicate whether a particular research area is more prone to QRPs than others. For instance, it may be interesting to compare strategy and nonstrategy research (e.g., human resource management and organizational behavior). Also, research on new (e.g., the effectiveness of mindfulness interventions) and established topics (e.g., general cognitive ability as a predictor of job performance) may differ in QRP usage. Fifth, a close examination of QRP occurrence and prevalence in individual journals could be worthwhile. Specifically, one could compare journals by the incidence of QRPs in their publications. For example, some editorial boards may encourage or be more tolerant of HARKing and other QRPs than other editorial boards. Our sixth recommendation is to consider methods that do not rely on comparisons of dissertations and articles. For instance, for the journal articles, one could code statistical parameters and run the data through statcheck (http://statcheck.io) and also use methods to check for statistical power and publication bias.

There are several important implications of our findings. Practitioners should be skeptical of findings published in our journal articles, especially in our most prestigious ones. Given the reward structures, our field should consider changing both department-level and journal-level policies and practices that motivate QRPs and instead develop policies that reward the behaviors that departments and journals likely desire—high-quality research. For instance, publishing standards for departments should be changed to not only reflect the number of articles published in journals with high impact factors but also with consideration to the quality, transparency, and methodological rigor of the articles. In addition, we should change the processes through which we implement and follow the scientific method. For decades, several researchers have explored and discussed problems in our research and publishing processes (e.g., Bedeian et al., 2010; Greenwald, 1975; Kepes & McDaniel, 2013; Rosenthal, 1979). Therefore, it seems to be time for our gatekeepers, especially our journal editors and leaders of our academic organizations, to implement some of the previously made recommendations, as described below. One recommendation in particular, the publishing of the comments from the editor and reviewers as well as the replies from the authors associated with every published article, perhaps as part of an article's

supplementary materials, could shed further light on the pressures emanating from our journals.

There have been several additional recommendations for improving the accuracy and trustworthiness of our scientific knowledge, ranging from the establishment of research registries (Kepes & McDaniel, 2013) and better data sharing requirements (Wicherts et al., 2006) to the publishing of exact replications (Makel et al., 2012) and constructive replications (Köhler & Cortina, 2019). We recommend that all journals follow the example of the APA journals (e.g., *Journal of Applied Psychology*) and endorse the Transparency and Openness Promotion (TOP) guidelines, which require transparency in several different areas as well as promote study preregistration. We also encourage alternative editorial review processes (Kepes et al., 2014b) that may also pay more attention to issues related to the statistical power of the submitted studies (Maxwell, 2004), such as the implementation of a triple blind review process.[9] In particular, the use of preregistration and results-blind review, such as the alternative submission process implemented by the *Journal of Business and Psychology*, may help to decrease the motivation to engage in QRPs. In this process, papers are evaluated based on their theory, hypotheses, method, and proposed data analysis plan (Journal of Business and Psychology, 2016; Kepes et al., 2014b). Overall, the objective of our sciences, to discover the truth about the world, is much better achieved with transparent scientific processes and an open research culture (Nosek et al., 2015). Finally, instead of judging the quality and prestige of our journals by their Impact Factor, which essentially assesses a journal's popularity, we should use metrics that assess the accuracy and trustworthiness of the research a journal publishes (such as the recently released TOP factor; Center for Open Science, 2020; Kepes et al., 2020).

5.2 | Conclusion

Environmental factors likely motivate researchers to engage in QRPs to enhance the probability that a paper will be published. We found that authors in our sample tend to suppress unsupported dissertation hypotheses by excluding them from the journal articles derived from the dissertations. This data suppression increases the percentage of supported hypotheses in the journal articles. In addition, our results are consistent with the inference that many article hypotheses were likely created after the results were known (i.e., HARKed hypotheses). Lastly, prestigious journals tend to contain a larger percentage of potentially HARKed article hypotheses when compared with less prestigious journals and the use of QRPs appears to be more common in articles published in high-prestige journals. Thus, although society hopes that universities, scientific journals, and researchers work in concert to generate accurate and credible scientific knowledge, environmental pressures (e.g., at the departmental and journal levels) may yield opposite outcomes, scientific reporting omissions and commissions. In fact, it seems as if departments and journals are "rewarding A, while hoping for B" (Kerr, 1975, p. 769).

ENDNOTES

[1] O'Boyle et al. (2017, p. 377) labeled outcome-reporting bias stemming from the use of QRP's the *Chrysalis Effect* "after the metamorphosis process whereby an ugly caterpillar (initial results) turns into a beautiful butterfly (journal article)."

[2] Some examples of nontrolled hypotheses include hypotheses added as a result of feedback by readers of the dissertation, including hypotheses in the article that were initially included as future research ideas in the dissertation and evolution of the author's perspective not based on the data analysis.

[3] Hypotheses H4b and H4c were added during the editorial review process (we thank the editor and the anonymous reviewers for this suggestion). We note that these hypotheses are in line with our original theorizing, which did not change when these hypotheses were added. Furthermore, we did not know the results of the analyses before adding these two hypotheses. Thus, they are not HARKed. We provide this information to be transparent.

[4] The originally submitted version of our paper looked at the top five management programs. Due to requests from the editor and reviewers, the sample was expanded to the top 10 programs during the editorial review process. The obtained results remained largely the same.

[5] We thank an anonymous reviewer for their comments which encouraged us to conduct these additional sensitivity analyses.

[6] We emphasize that hypothesis suppression can take place prior to submitting a manuscript to a journal or in response to a request during the editorial review process.

[7] We used more than three decimal places in calculating the risk ratios (0.1617647 / 0.09767442 rounds to 1.66).

[8] An anonymous reviewer suggested that dissertations may contain both "core" and "peripheral" hypotheses and that "core" hypotheses may be more carefully developed and thus more likely to be associated with statistically significant results. When authors turn their dissertations into articles, they drop the "peripheral" hypotheses (which also happen to be unsupported at a higher rate than the "core" hypotheses). In this scenario, the systematic dropping of "peripheral" hypotheses would look similar to a situation where authors are dropping hypotheses based on whether they are associated with statistically significant results. Therefore, though the idea about "core" and "peripheral" hypotheses may have merit, it is unlikely to change the dynamics—statistically nonsignificant results would still be systematically suppressed from the publicly available literature. Also, it is likely that "peripheral" dissertation hypotheses that are supported will be part of a journal article—either with the core hypotheses in one article or in a separate one. That may partly explain why many dissertations yield more than one article; one article includes the "core" hypotheses (and associated results) and other article(s) the "peripheral" hypotheses (and associated results).

[9] We thank an anonymous reviewer for this suggestion.

REFERENCES

Agnew, R. (1992). Foundation for a general strain theory of crime and delinquency. *Criminology*, *30*, 47–88. https://doi.org/10.1111/j.1745-9125.1992.tb01093.x

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, *436*, 900. https://doi.org/10.1038/436900a

Banks, G. C., Field, J. G., Oswald, F. L., O'Boyle, E. H., Landis, R. S., Rupp, D. E., & Rogelberg, S. G. (2019). Answers to 18 questions about open science practices. *Journal of Business and Psychology*, *34*, 257–270. https://doi.org/10.1007/s10869-018-9547-8

Banks, G. C., Kepes, S., & McDaniel, M. A. (2015). Publication bias: Understanding the myths concerning threats to the advancement of science. In C. E. Lance & R. J. Vandenberg (Eds.), *More Statistical and Methodological Myths and Urban Legends* (pp. 36–64). Routledge.

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., & Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, *42*, 5–20. https://doi.org/10.1177/0149206315619011

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, *31*, 1–16. https://doi.org/10.1007/s10869-016-9456-7

Becker, T. E. (2005). Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, *8*, 274–289. https://doi.org/10.1177/1094428105278021

Bedeian, A. G., Taylor, S. G., & Miller, A. N. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education*, *9*, 715–725.

Bloom, M. (1999). The performance effects of pay dispersion on individuals and organizations. *Academy of Management Journal*, *42*, 25–40.

Campbell, J. P., & Wilmot, M. P. (2018). The functioning of theory in industrial, work and organizational psychology (IWOP). In *The SAGE Handbook of Industrial, Work & Organizational Psychology: Personnel Psychology and Employee Performance* (2nd ed., Vol. 2, pp. 3–38). Sage Reference. https://doi.org/10.4135/9781473914940.n2

Carlson, K. D., & Wu, J. (2011). The illusion of statistical control: Control variable practice in management research. *Organizational Research Methods*, *15*, 413–435. https://doi.org/10.1177/1094428111428817

Center for Open Science. (2020). New measure rates quality of research journals' policies to promote transparency and reproducibility. Retrieved February 24, 2020, from https://cos.io/about/news/new-measure-rates-quality-research-journals-policies-promotetransparency-and-reproducibility/

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*, 867–872. https://doi.org/10.1016/0140-6736(91)90201-Y

Editors. (2006). Cash-per-publication. In *Nature* (Vol. 441, p. 786). https://doi.org/10.1038/441786a

Estes, B., & Polnick, B. (2012). Examining motivation theory in higher education: An expectancy theory analysis of tenured faculty productivity. *International Journal of Management, Business, and Administration*, *15*, 1–7.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904. https://doi.org/10.1007/s11192-011-0494-7

Franco, A., Malhotra, N., & Simonovits, G. (2014). Social science. Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*, 1502–1505. https://doi.org/10.1126/science.1255484

Gomez-Mejia, L. R., & Balkin, D. B. (1992). Determinants of faculty pay: An agency theory perspective. *Academy of Management Journal*, *35*, 921–955.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20. https://doi.org/10.1037/h0076157

Grote, G., & Cortina, J. M. (2018). Necessity (not just novelty) is the mother of invention: Using creativity research to improve research in work and organizational psychology. *European Journal of Work and Organizational Psychology*, *27*, 335–341. https://doi.org/10.1080/1359432X.2018.1444606

Haensly, P. J., Hodges, P. E., & Davenport, S. A. (2008). *Acceptance Rates and Journal Quality: An Analysis of Journals in Economics and Finance*. Journal of Business & Finance Librarianship, *14*, 2–31. https://doi.org/10.1080/08963560802176330

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. https://doi.org/10.1177/0956797611430953

Journal of Business and Psychology. (2016). JBP special initiative results-blind review resource site. Retrieved August 13, 2021, from https://jbp.uncc.edu/.10.31957/jbp.1640

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. https://doi.org/10.2307/1914185

Kemper, T. D. (1978). *A Social Interactional Theory of Emotions*. Wiley.

Kepes, S., Banks, G., McDaniel, C, M, A., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, *15*, 624–662. https://doi.org/10.1177/1094428112452760

Kepes, S., Banks, G. C., & Keener, S. K. (2020). The TOP factor: An indicator of quality to complement journal impact factor. *Industrial and Organizational Psychology*, *13*, 328–333. https://doi.org/10.1017/iop.2020.56

Kepes, S., Banks, G. C., & Oh, I.-S. (2014a). Avoiding bias in publication bias research: The value of. *"Null" Findings. Journal of Business and Psychology*, *29*, 183–203. https://doi.org/10.1007/s10869-012-9279-0

Kepes, S., Bennett, A. A., & McDaniel, M. A. (2014b). Evidence-based management and the trustworthiness of our cumulative scientific knowledge: Implications for teaching, research, and practice. *Academy of Management Learning & Education*, *13*, 446–466. https://doi.org/10.5465/amle.2013.0193

Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *6*, 252–268. https://doi.org/10.1111/iops.12045

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

Kerr, S. (1975). On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, *18*, 769–783.

Köhler, T., & Cortina, J. M. (2019). Play it again, Sam! An analysis of constructive replication in the organizational sciences. *Journal of Management*, *47*, 488–518. https://doi.org/10.1177/0149206319843985

Lawler, E. E. (1971). *Pay and Organizational Effectiveness: A Psychological View*. McGraw-Hill.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542. https://doi.org/10.1177/1745691612460688

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies.

Psychological Methods, *9*, 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*, *34*, 1–17. https://doi.org/10.1007/s10869-017-9524-7

Murtaugh, P. A. (2002). Journal quality, effect size, and publication bias in meta-analysis. *Ecology*, *83*, 1162–1166. https://doi.org/10.1890/0012-9658(2002)083[1162:JQESAP]2.0.CO;2

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425. https://doi.org/10.1126/science.aab2374

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*, 615–631. https://doi.org/10.1177/1745691612459058

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mule, E. (2017). The Chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, *43*, 376–399. https://doi.org/10.1177/0149206314527133

Ostriker, J. P., Holland, P. W., Kuh, C. V., & Voytuk, J. A. (2009). *A Guide to the Methodology of the National Research Council Assessment of Doctorate Programs*. The National Academies Press.

Petit, N. C., Yong, K., & Spataro, S. E. (2010). Holding your place: Reactions to the prospect of status gains and losses. *Journal of Experimental Social Psychology*, *46*, 396–401. https://doi.org/10.1016/j.jesp.2009.12.007

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*, 424–432. https://doi.org/10.3102/0013189X13507104

Podsakoff, P. M., MacKenzie, S. B., Podsakoff, N. P., & Bachrach, D. G. (2008). Scholarly influence in the field of management: A bibliometric analysis of the determinants of university and author impact in the management literature in the past quarter century. *Journal of Management*, *34*, 641–720. https://doi.org/10.1177/0149206308319533

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rupp, D. E. (2011). Ethical issues faced by editors and reviewers. *Management and Organization Review*, *7*, 481–493. https://doi.org/10.1111/j.1740-8784.2011.00227.x

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Sterling, T. D., & Rosenbaum, W. L. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician*, *49*, 108–112.

Tosi, H. L., Werner, S., Katz, J. P., & Gomez-Mejia, L. R. (2000). How much does performance matter? A meta-analysis of CEO pay studies. *Journal of Management*, *26*, 301–339. https://doi.org/10.1177/014920630002600207

Vroom, V. H. (1964). *Work and Motivation*. Wiley.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, *6*, e26828.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728. https://doi.org/10.1037/0003-066X.61.7.726

AUTHOR BIOGRAPHIES

Sven Kepes is a Professor of Management at Virginia Commonwealth University. His research interests include strategic human resource management, compensation systems, meta-analytic methods, publication bias, research integrity, and the trustworthiness of our cumulative knowledge. His research has been published in journals such as Psychological Bulletin, Journal of Applied Psychology, and Organizational Research Methods.

Sheila K. Keener is an Assistant Professor of Management at Old Dominion University. Her main research interests are in the areas of Human Resources Management and improving the quality and integrity of research. Her research has been published in journals such as the Journal of Applied Psychology and Organizational Research Methods.

Michael A. McDaniel is an Emeritus Professor of Management at Virginia Commonwealth University. His research focused on personnel selection, the application of meta-analysis to management research areas, publication bias, and research integrity. His research has been published in several journals, including the Journal of Applied Psychology, Personnel Psychology, and Organizational Research Methods.

Nathan S. Hartman is an Associate Professor at Illinois State University. His primary research interests include organizational citizenship behaviors, employee selection, and leader development. He has authored articles in publications including the Journal of Management Development, Organizational Research Methods, Group & Organization Management, and Personnel Psychology.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.