Journal of Management
Vol. XX No. X, Month XXXX 1–24
DOI: 10.1177/0149206314527133

The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles

Ernest Hugh O'Boyle Jr.
University of Iowa
George Christopher Banks
Longwood University
Erik Gonzalez-Mulé
University of Iowa

The issue of a published literature not representative of the population of research is most often discussed in terms of entire studies being suppressed. However, alternative sources of publication bias are questionable research practices (QRPs) that entail post hoc alterations of hypotheses to support data or post hoc alterations of data to support hypotheses. Using general strain theory as an explanatory framework, we outline the means, motives, and opportunities for researchers to better their chances of publication independent of rigor and relevance. We then assess the frequency of QRPs in management research by tracking differences between dissertations and their resulting journal publications. Our primary finding is that from dissertation to journal article, the ratio of supported to unsupported hypotheses more than doubled (0.82 to 1.00 versus 1.94 to 1.00). The rise in predictive accuracy resulted from the dropping of statistically nonsignificant hypotheses, the addition of statistically significant hypotheses, the reversing of predicted direction of hypotheses, and alterations to data. We conclude with recommendations to help mitigate the problem of an unrepresentative literature that we label the "Chrysalis Effect."

Keywords: *philosophy of science; statistical methods; ethics; morality and moral behavior*

Acknowledgments: An earlier version of this manuscript was presented at the 2013 Academy of Management Conference in Orlando, Florida. The authors would like to acknowledge Ken Brown, Eean Crawford, Sven Kepes, Ning Li, Mick Mount, In-Sue Oh, and Frank Schmidt for their helpful feedback in preparing the manuscript for submission. The authors extend a special thanks to Mike McDaniel for his 2012 presentation for the Center for the Advancement of Research Methods and Analysis that helped serve as the impetus for this work. Finally, we wish to thank Fred Oswald and two anonymous reviewers for their support and thoughtful feedback during the revision process.

Corresponding author: Ernest Hugh O'Boyle Jr., University of Iowa, W332 Pappajohn Business Building, Iowa City, IA 52242-1994, USA.

E-mail: oboyleeh@gmail.com

Publications in refereed journals, particularly, top-tier journals, are the currency in which our field trades and are a crucial antecedent to meaningful outcomes, such as salary and tenure (Gomez-Mejia & Balkin, 1992). A reward structure that relies primarily on journal publications ensures the continuing research productivity of those entering and establishing themselves in the field of management. Yet, coupled with other characteristics of the field, such a reward structure also lends itself to an intense pressure to publish. These pressures may lead researchers to engage in questionable research practices (QRPs) to bolster their chances of publication (Chen, 2011; Fanelli, 2010a, Kepes & McDaniel, 2013). Unlike blatant scientific misconduct (e.g., fabricating data), QRPs skirt the line between ethical and unethical behavior. Examples of QRPs include presenting post hoc findings as a priori hypotheses and dropping data points based on post hoc criteria to achieve statistically significant results (Leung, 2011). The result of these QRPs is an increase in Type I errors and a suppression of null effects, which biases the literature (John, Loewenstein, & Prelec, 2012). We label this form of outcome-reporting bias the "Chrysalis Effect" after the metamorphosis process whereby an ugly caterpillar (initial results) turns into a beautiful butterfly (journal article).

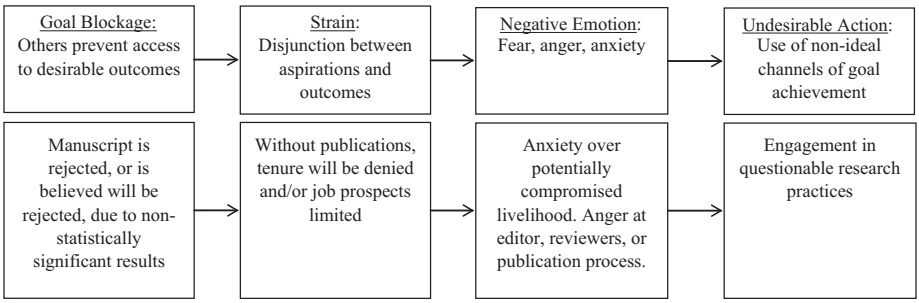
Applying the sociological perspective of general strain theory (Agnew, 1992; Cloward & Ohlin, 1960; A. Cohen, 1955; Merton, 1938), we outline the means, motives, and opportunity of improving the likelihood of publication independent of rigor and relevance. We then track changes in 142 dissertations that were subsequently published in refereed journals. By examining the characteristics of dropped and added hypotheses, fluctuations in sample size, and alterations to data, we are able to assess the degree to which QRPs affect the stability and veracity of the extant literature. We conclude with a series of recommendations for researchers that would help remedy the problems associated with the Chrysalis Effect.

The Chrysalis Effect and General Strain Theory

A number of commentaries have addressed QRPs (e.g., John et al., 2012; Kacmar, 2009; Schminke, 2009), but it is difficult to determine the presence and extent of the problem through self or peer reports given concerns over social desirability and nonresponse bias. Although their prevalence cannot fully be determined, we argue that attributing these behaviors to a few bad apples both understates the problem and misattributes the cause to the individual, when it is most likely systemic. As such, we believe the sociological perspective of general strain theory is an appropriate framework for explaining how and why the Chrysalis Effect occurs. Not only does the theory map well onto the motivations and practices to be discussed; the fact that it is a sociological theory further reinforces our contention that the cause of QRPs are likely to be the system, not the person.

General strain theory views undesirable behavior from the perspective of negative social relationships that can be defined as "any relationship in which others are not treating the individual as he or she would like to be treated" (Agnew, 1992, p. 50). It has most often been applied to crime and delinquency (Agnew, Brezina, Wright, & Cullen, 2002), but we contend that it is just as applicable to less severe behaviors, such as QRPs. Figure 1 provides an overview of general strain theory and how it can foster non-ideal publication practices. According to the theory, when others prevent access, or are believed will prevent access, to the desirable goal of publication, the researcher experiences strain caused by the disjunction of his or her aspiration and the reality of the outcome (i.e., failure to publish).

Figure 1
General Strain Theory Applied to Questionable Research Practices



Worth noting is that the blocking of desirable outcomes and subsequent strain is caused by a negative social relationship, but it is not necessarily a malevolent relationship. For example, editors and reviewers routinely block authors' work from being published, but the peer review process is an integral and largely positive aspect of scientific research. Nevertheless, the strain creates negative emotional states (e.g., anger at reviewers, anxiety over publication record) that in certain high-stakes situations (e.g., a looming tenure clock) motivate undesirable behavior (Agnew, 1992; Merton, 1938). Germane to the Chrysalis Effect is that the strain creates discomfort (similar to that of cognitive dissonance) that a person might attempt to remedy by use of non-ideal channels of goal achievement (Kemper, 1978). If an author believes that the initial results of a study are insufficient for publication, he or she may attempt to achieve publication by way of QRPs.

We argue that the publication process and reward systems in academia provide the means, motives, and opportunity to engage in QRPs. Specifically, empirical research has documented the continued implicit reliance in peer review systems on null hypothesis significance testing as a proxy for rigor and relevance (e.g., Emerson et al., 2008). Yet contrary to what is desired, the dependence on such a criterion offers researchers one set of means to publication independent of rigor and relevance. In addition, the compensation system found in academia is a tournament-style reward system (Rynes & Gerhart, 2003), which places great emphasis on publications. This is a largely positive system that motivates innovation and new management theory, but it also creates a motive to engage in QRPs. Finally, the lack of oversight in the research and review process provides the opportunity to engage in QRPs with little chance of detection.

Means of QRPs

If science were a game and publications the prize, then the "dominant rule would probably be to collect results that are statistically significant" (Bakker, van Dijk, & Wicherts, 2012, p. 543). A review of the literature suggests that as a field, we are winning the game despite management research often being statistically underpowered with relatively small samples and effect sizes (Cashen & Geiger, 2004; J. Cohen, 1990; Marszalek, Barber, Kohlhart, & Holmes, 2011). There has been a repeated call for management researchers, reviewers, and

editors to stop overemphasizing statistical significance in their evaluation of research quality (J. Cohen, 1994; Gigerenzer, 2004; Meehl, 1978; Nickerson, 2000; Schmidt, 1992, 1996; Schmidt & Hunter, 2002; Wagenmakers, 2007). Despite this, most researchers assume that statistically significant results are more likely to be published than nonsignificant results (Bakker et al., 2012), and there is evidence that this assumption is not entirely unfounded (Orlitzky, 2012).

If the perception is that statistically nonsignificant results are less likely to be published, then one set of means by which a researcher achieves publication through non-ideal channels is by altering hypotheses to support statistically significant data or by altering nonstatistically significant data to support hypotheses. Below, we review five QRPs that researchers may be using to increase their likelihood of publication. These five are by no means an exhaustive list, but we focus on these QRPs because most empirical management research still relies on a priori hypotheses of which acceptance or rejection is largely determined by a *p* value.

Deletion or addition of data after hypothesis tests. When hypotheses are not supported with statistical significance, especially when they are nearing the significance threshold, there may be a temptation for researchers to either add cases to increase statistical power or return to the data looking to delete cases that raise *p* values above the .05 threshold (Aguinis, Gottfredson, & Joo, in press). To add or drop cases because results are nearly supportive of hypotheses (e.g., $.05 < p < .10$) defeats the purpose of a priori statistical significance testing. Nevertheless, there is evidence to suggest this practice occurs in related fields, such as psychology, sociology, and political science, where these literatures show an abnormally high prevalence of *p* values just below the .05 threshold (Gerber & Malhotra, 2008; Masicampo & Lalande, 2012).

Altering the data after hypothesis testing. The Chrysalis Effect may also manifest itself via data distortion. When done post hoc for the purposes of supporting hypotheses, practices such as dropping items from scales or categorizing a continuous variable results in an inflation of Type I errors. Although it is difficult to determine whether data were specifically altered to support hypotheses, we can document changes in the means, variances, and relations to other variables and whether these changes were associated with an increase in the number of hypotheses supported by statistical significance.

Selective deletion or addition of variables. In theory-driven research, the inclusion of a variable, even a control variable, is believed to be somehow important to the research questions. To drop variables from a research project with no explanation is suspect, and the ability to do so with impunity encourages a kitchen-sink approach to research (Chalmers, 1990). Similarly, adding variables to a research effort after hypotheses are tested can be equally counterproductive when their addition is data-driven instead of theory-driven. As Mills (1993) noted, "If you torture your data long enough, they will tell you whatever you want to hear" (p. 1196).

Reversing the direction or reframing hypotheses to support data. It is not uncommon, especially with complex hypotheses (e.g., three-way interactions, nonlinear relations), for results to be statistically significant but in the opposite direction as predicted. Insight can still be gained from these surprising findings, hence the convention of two-tailed statistical

significance tests, but if the a priori hypothesis was not supported, then changing the hypothesis to conform to the data capitalizes on chance (Bedeian, Taylor, & Miller, 2010; Kerr, 1998). Once again quoting Mills (1993), "if the fishing expedition catches a boot, the fishermen should throw it back, not claim that they were fishing for boots" (p. 1198).

Post hoc dropping or adding of hypotheses. The hypothetico-deductive model that most management research operates in dictates that hypotheses are made before data are analyzed. Contrary to this tenet, researchers may drop hypotheses that are either not statistically significant or contrary to the expectations of the researcher or add hypotheses after initial results are known (John et al., 2012; Kerr, 1998; Leung, 2011).

QRPs and the Chrysalis Effect. It is worth noting that these five QRPs bias the literature only if they alter the ratio of supported to unsupported hypotheses. For example, there are many reasons to add and drop hypotheses that would not generate a Chrysalis Effect, such as a reviewer suggesting the addition of an interaction hypothesis or one of the measures used to test the hypothesis exhibiting unacceptable psychometric properties. However, if the adding and dropping of hypotheses were based purely on rigor and relevance, then the ratio of statistically significant to nonsignificant dropped hypotheses should be approximately equal to the ratio of statistically significant to nonsignificant added hypotheses. If there was a large disparity between the two ratios (i.e., dropped hypotheses were typically unsupported and added hypotheses were typically supported), then the practice of adding or deleting hypotheses is likely motivated by statistical significance.

Motives for QRPs

We identify two critical motives to engage in QRPs. However, both motives serve the same purpose of addressing the strain caused by the disjunction between aspirations and outcomes. That is, the motive is to reduce strain, and the means to do so is through QRPs.

Anticipation of reviewer reactions or compliance to reviewer requests. Ironically, pressure to engage in QRPs that harm the field can come from the very gatekeepers responsible for preserving its integrity. The peer review system most often makes manuscripts better in terms of rigor and relevance, and the preponderance of changes that occur during this process are positive. However, the well-intentioned developmental review can lead researchers to make post hoc alterations to theory and hypotheses based on the recommendations of editors and reviewers (Bedeian, 2004). For example, a survey of authors whose work was published in *Academy of Management Journal* and *Academy of Management Review* found that nearly one in four made an editor- or reviewer-directed change to the manuscript they knew was incorrect (Bedeian, 2003). Similarly, and likely a more common occurrence, is that authors make changes to their manuscript before submission in anticipation of reviewer reactions (Bakker et al., 2012). If either perceived reviewer reactions or actual review requests lead to changes that are more likely directed at statistically nonsignificant hypotheses than their significant counterparts, then these changes gild the published literature by giving the appearance of theoretical grounding when in fact many of the changes were post hoc and data driven.

Journal lists and publication heuristics. Beginning in the early 1990s, there was a noticeable increase in business schools trying to determine what constitutes a top-tier publication (Corley & Gioia, 2000). The pressure to formalize and quantify journal quality typically came from top administration whose concern was that certain journals were incorporated into business school ranking systems, such as *Business Week* and *US News & World Report*, and others were not (Gioia & Corley, 2002; Siemens, Burton, Jensen, & Mendoza, 2005). The creation of journal lists narrowed the acceptable or desirable outlets for management research. For some researchers, journal lists make goal achievement more difficult and may generate additional strain, leading to QRPs. Further, our field's reward system, via citations, encourages the publication of papers that do not depend on primary data (e.g., meta-analyses, simulation studies, and reviews). Therefore, to compete with these types of papers, scholars may feel they need to engage in QRPs to increase their perceived chances of gaining a top-tier publication.

Beyond quality, there is also the question of quantity. Selection as well as promotion and tenure committees may rely in part on a minimum number of needed publications (Buchheit, Collins, & Reitenga, 2002; MacDonald & Kam, 2007). On the one hand, this provides doctoral students and assistant professors with specific and measurable goals, and tenure clocks make these goals time bound. Goal-setting theory (Locke & Latham, 1990) would suggest that this increases the research output of both doctoral candidates and junior faculty. On the other hand, when there is goal blockage, the importance of the outcome (i.e., their livelihood) creates the strain and subsequent negative emotions that might lead one to publish via QRPs.

Opportunity for QRPs

Regardless of the means and motives to better one's chances of publication through non-ideal practices, it is unlikely that researchers would engage in QRPs if the probability of detection were high. Thus, the near-complete lack of oversight in data collection and analysis provides management researchers the opportunity to engage in QRPs with little fear of detection.

Data privacy and reporting practices. Despite the significant concern placed upon Type I errors, "flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates" (Simmons, Nelson, & Simonsohn, 2011, p. 1359). The current state of the sciences is that most researchers are unwilling to share data (Ioannidis, 2005, 2008; Nicholas & Katz, 1985), and authors are given high degrees of latitude in what they report (Klein, Doyen, Leys, Miller, Questienne, & Cleeremans, 2012). Limited access to the original data and insufficient reporting standards create conditions whereby QRPs are virtually undetectable.

Lack of replication. Replication serves as the principal means by which the reliability of a finding can be established (Frank & Saxe, 2012). Replication can occur when authors make their data available for reanalysis, but as described above, this is exceedingly rare (Ioannidis, 2005, 2008). Researchers can also conduct a replication study with different subjects, but this is also rare in management science because there is little incentive to do so (Eden, 2002; Rosenthal, 1990). If the results replicate, then there is a large chance that the paper will be dismissed as "old news." If the results do not replicate, then the author faces the uphill battle of justifying the importance of null findings.

The Present Study

We tested for the Chrysalis Effect in management by tracking changes in hypotheses, data, and results as a manuscript moved from defended dissertation to journal publication. We chose this process for several reasons. First, although a researcher's career contains multiple points of strain depending on his or her unique situation (e.g., cusp of tenure, move to a more research-oriented institution, remaining academically qualified), publishing early in one's career as an untenured assistant professor is nearly a universal point of great strain. Second, although there is debate about the extent that statistically significant findings are favored over statistically nonsignificant findings in the journal publication process (e.g., Dalton, Aguinis, Dalton, Bosco, & Pierce, 2012), there is less concern that this favoritism extends to the dissertation process, where proposals are typically offered before data are collected and analyzed. Third, the dissertation is closely supervised by senior faculty members, making QRPs less likely to go undetected.

Method

Identification of Studies and Inclusion Criteria

Our interest was in changes at both the hypothesis level and the project level (i.e., the dissertation and any resulting publication(s) from the dissertation). We used ProQuest Dissertations and Theses to identify our pool of potential studies. ProQuest contains dissertations from both the physical and social sciences, and we wished to limit the search to doctoral dissertations with a management-relevant topic. As such, we used search terms that we believed would result in dissertations with a largely management focus (e.g., *workplace deviance, leader-member exchange, resource based view, entrepreneurial orientation*), and we excluded dissertations that either were not from management or applied psychology departments or were not published in management or applied psychology journals. We also excluded qualitative research, case studies, systematic reviews (i.e., meta-analyses), and dissertations that lacked formal hypotheses. Because we were most interested in current practices, we limited our search to dissertations published between 2000 and 2012. We believed this range relevant, as it is likely to contain the largest segment of those actively seeking tenure, which as stated above, is a common point of great strain.

After identifying a dissertation on a management topic, we then entered the search term and the author's name into Google Scholar and searched for any resulting publication(s) based upon the dissertation. This proved to be more difficult than expected because (a) the title often changed from dissertation to publication, and (b) only 18% of the articles we identified as based upon a dissertation acknowledged the prior work. Regarding the latter point, the low percentage of those acknowledging their dissertations foreshadowed what we were to find. Dissertations are copyrighted and the ethical guidelines of our governing agencies (i.e., Academy of Management, 2006, Code of Ethics Section 4.1.2; American Psychological Association, 2010, Ethics Code Section 8.13) require acknowledgement that data were previously published. Failure to do so is yet another QRP.

We then used Wood's (2008) detection heuristics to determine if there was overwhelming evidence that the journal publication was in fact based on the research conducted for the dissertation. When in doubt, we eliminated the study from further consideration.

We repeated this process for the first 100 results (sorted by relevance) of each search term. We limited ourselves to the first 100 results as our experience was that except for the most popular topics (e.g., agency theory), the dissertations beyond the first 100 tended to be tangentially related to the search term and rarely included a test of the theory or a measure of the construct. Although this limited our potential pool of studies to slightly more than 2,000, we were still able to identify 142 dissertations where there was overwhelming evidence that it had been subsequently published in a refereed journal.

It should be noted that this search process diverges from a traditional systematic review (e.g., meta-analysis) where keywords and search criteria target a focal construct or relationship and seek to identify the entirety of research on a given topic. We wished to achieve breadth in our search more so than depth on QRPs associated with a specific construct or relationship. Because we did not identify the entire pool of dissertations, this potentially limits the generalizability of our results to the population of dissertations that were subsequently published in journals. However, we do not expect that the topics we selected inherently bias the frequency or effects of QRPs.

Coding of Studies

For both the dissertations and journal articles, we coded (a) type of study (field or experiment), (b) date of publication, (c) included variables, (d) sample sizes, (e) whether a committee member was a coauthor on the article, (f) author affiliation at time of journal publication, (g) hypotheses, (h) changes in descriptive statistics and correlations, and (i) whether the hypotheses were statistically significant. We coded only hypotheses that were explicitly stated, but we did note when an exploratory research question or proposition in the dissertation was formalized to a directional hypothesis in the resulting publication. We also coded journal quality with Hirsch's (2005) *h*-index, which has been shown to provide more accurate measures of influence than commonly used metrics, such as the ISI Web of Knowledge impact factors (Harzing & van der Wal, 2008). Two of the authors independently coded a subsample of the included dissertations and journal publications, and across 120 coding decisions (e.g., hypothesis direction, statistical significance, adding of variables), the interrater reliability was adequate with a Cohen's kappa of .94 (J. Cohen, 1960). Any discrepancies in coding were resolved through discussion.

Analysis Strategy

We examined hypotheses in terms of their addition, deletion, or change in statistical significance. Because not all alterations to data or hypotheses are done so out of a desire to increase statistical significance or suppress nonstatistically significant findings, alterations alone neither support nor reject the Chrysalis Effect. However, if the Chrysalis Effect were present, then we would expect to see the ratio of supported to unsupported hypotheses increase from dissertation to journal publication. This would occur if added hypotheses were more likely to be statistically significant, dropped hypotheses were less likely to be statistically significant, and retained hypotheses (i.e., those common to both dissertation and journal) more often changed in statistical significance (Δ_{signif}) from unsupported to supported than vice versa.

Results

There were 89 schools represented and of these, 47 (53%) were institutions rated as "very high research activity" by the Carnegie Classification of Institutions of Higher Education. The most well-represented schools were University of Maryland (9), Florida State University (5), University of Minnesota (4), Ohio State University (3), Stanford University (3), Texas A&M University (3), and University of Kansas (3). Thirteen schools were located outside of the United States, with the majority (10) hailing from Canada. For the journal publications, there were 83 journals represented and the most frequent were *Journal of Applied Psychology* (10), *Journal of Organizational Behavior* (7), *Journal of Leadership & Organizational Studies* (7), and *Academy of Management Journal* (6). Other notable inclusions were *Human Resource Management* (4), *Organization Science* (3), *Journal of Management* (3), *Leadership Quarterly* (3), *Strategic Management Journal* (2), *Personnel Psychology* (2), and *Administrative Science Quarterly* (1). The average date of the dissertations was 2006 (*SD* = 2.76 years), and the average date of journal publication was 2009 (*SD* = 2.39 years). The average time to publication was 3.39 years (*SD* = 1.89 years).

There were 149 field studies and 26 experiments (the value exceeds 142 because of multistudy projects). For each QRP in Tables 1 and 2, we present the corresponding hypotheses that appeared in both the dissertation and the published journal article (i.e., common hypotheses) and the hypotheses that were unique to either the dissertation or the journal article. Tables 1 and 2 also provide the percentage differences and risk ratios (RRs) of the various QRPs. We chose RRs over odds ratios because of the persistent confusion in the interpretation of odds ratios that often results in overestimating effect sizes (Bracken & Sinclair, 1998; Davies, Crombie, & Tavakoli, 1998; Knol, Duijnhoven, Grobbee, Moons, & Groenwold, 2011). For example, an RR of 3.0 indicates that an event such as a hypothesis being statistically significant is three times as likely to occur under one condition (e.g., added to a journal article) than another condition (e.g., dropped from a dissertation). An odds ratio can have a very different interpretation, especially when events occur at frequencies greater than 10% (D. Sackett, Deeks, & Altman, 1996).

Across both the dissertations and journal articles, we recorded 2,311 hypotheses. The dissertations tested 1,978 hypotheses and their corresponding publications tested 978 hypotheses, a net change of 1,000. There were 645 common or retained hypotheses that remained essentially unchanged from dissertation to journal article. Of these common hypotheses, 373 were supported in the dissertation (57.8%, a ratio of supported to unsupported hypotheses of 1.37:1), and 412 were supported in the journal article (63.9%, a ratio of supported to unsupported hypotheses of 1.77:1). The remaining 1,666 hypotheses were either dropped from the published article or added after the dissertation defense.

We next document the prevalence of each of the QRPs and their relations to the statistical significance of hypotheses. We then compare the ratio of supported to unsupported hypotheses in the dissertations (common and dropped hypotheses) to the ratio of supported to unsupported hypotheses in the journal articles (common and added hypotheses). The discrepancy between the two is the estimate of the overall Chrysalis Effect.

Table 1
Summary of QRP Engagement Among Hypotheses Common to Both Dissertation and Journal Article

	Unsupported Dissertation Hypothesis			Supported Dissertation Hypothesis				
QRPs and Common Hypotheses	*n*	$\Delta_{support}$	Percentage	*N*	$\Delta_{support}$	Percentage	% Diff.	Risk Ratio (95% CI)
Total[a]	272	56	20.6	373	17	4.6	16.0	4.52*** [2.69, 7.60]
Add/drop data	53	13	24.5	108	11	10.2	14.3	2.35* [1.16, 5.01]
Add data	21	4	19.0	45	4	8.9	10.1	2.14 [0.59, 7.75]
Drop data	32	9	24.5	63	7	11.1	13.4	2.53* [1.04, 6.17]
Alter data	47	16	34.0	63	0	0.0	34.0	N/A
Add variables	84	25	29.8	136	11	8.1	21.7	3.68*** [1.91, 7.08]
Drop variables	146	36	24.7	215	15	7.0	17.7	3.53*** [2.01, 6.22]
Altered hypothesis	22	17	77.3	0	0	—	—	N/A

Note: QRP = questionable research practice. The upper portion of the table divides common hypotheses (i.e., appeared in both dissertation and publication) into those that were initially supported with statistical significance from those that were not. $\Delta_{support}$ and percentage are the number and percentage of hypotheses that changed in statistical significance from dissertation to journal. Risk ratio in the upper portion is the likelihood of a hypothesis changing from nonstatistically significant to significant compared to the likelihood that it changed from statistically significant to nonsignificant. Risk ratio in the lower portion of the table is the likelihood that an added hypothesis was significant compared to the likelihood that a dropped hypothesis was significant. % Diff. in the upper portion is the percentage difference between unsupported and supported hypotheses changing in statistical significance. % Diff. in the lower portion is the percent difference between added and dropped hypotheses. In both the upper and lower portions, *n* denotes the number of hypotheses. Signif. = statistically significant at the .05 level; N/A = not applicable due to zero events; 95% CI = 95% confidence interval around the risk ratio.
[a]Sums of individual categories are greater than the total due to QRP overlap.
*$p < .05$.
**$p < .01$.
***$p < .001$.

Table 2
Summary of QRP Engagement Among Hypotheses Unique to Dissertation or Journal Article

QRPs and Unique Hypotheses	*n*	Signif.	Percentage	% Diff.	Risk Ratio [95% CI]
Added hypothesis	333	233	70.0		
Dropped hypothesis	1,333	516	38.7	31.3	1.81*** [1.64, 1.99]

Note: QRP = questionable research practice. Risk ratio is the likelihood that an added hypothesis was significant compared to the likelihood that a dropped hypothesis was significant. % Diff. is the percent difference between added and dropped hypotheses. *N* denotes the number of hypotheses. Signif. = number of hypotheses that were statistically significant at the .05 level; 95% CI = 95% confidence interval around the risk ratio.
***$p < .001$.

QRPs Associated With Common (Retained) Hypotheses

If the Chrysalis Effect were present, then changes to data, variables, and boundary conditions would more likely affect the 272 unsupported dissertation hypotheses than the 373

supported dissertation hypotheses. That is, it would be more likely for an unsupported dissertation hypothesis to become a supported journal hypothesis than a supported dissertation hypothesis to become an unsupported journal hypothesis. As Table 1 shows, this was the case. Among the dissertation hypotheses not supported with statistical significance, 56 of 272 (20.6%) turned into statistically significant journal hypotheses as compared to 17 of 373 (4.6%) supported dissertation hypotheses becoming statistically nonsignificant journal hypotheses. As indicated by the RR of 4.52 (95% CI = [2.69, 7.60]), hypotheses not supported with statistical significance were more than four times as likely to become supported hypotheses than vice versa. Below, we outline the specific QRPs associated with these changes.

Deletion or addition of data after hypothesis tests. Across the 142 projects, 14 (9.9%) added subjects (as evidenced by increases in sample size from dissertation to journal) and 29 (20.4%) dropped subjects (as evidenced by decreases in sample size from dissertation to journal). Within these 43 studies where sample size increased or decreased, there were 53 unsupported (i.e., statistically nonsignificant) and 108 supported (i.e., statistically significant) dissertation hypotheses retained in the journal publication. Thirteen of the 53 (24.5%) unsupported hypotheses became supported publication hypotheses as compared to only 11 of 108 (10.2%) supported dissertation hypotheses changing to statistically nonsignificant. In sum, changes in sample sizes were more than twice as likely to result in unsupported dissertation hypotheses changing to statistically significant journal hypotheses than vice versa (RR = 2.35; 95% CI = [1.16, 5.01]).

When divided into those that added subjects and those that dropped subjects, the results were similar. Among the studies that added subjects, 4 of 21 (19.0%) unsupported dissertation hypotheses changed to statistically significant in the journal article as compared to 4 of 45 (8.9%) supported dissertation hypotheses becoming statistically nonsignificant (RR = 2.14; 95% CI = [0.59, 7.75]). Because the confidence interval includes 1.00, this difference is not statistically significant at the .05 level. Among the studies that dropped subjects, 9 of 32 (28.1%) hypotheses became statistically significant from dissertation defense to journal publication, and 7 of 63 (11.1%) supported hypotheses in the dissertation became unsupported in the journal article (RR = 2.53; 95% CI = [1.04, 6.17]).

Altering the data after hypothesis testing. We examined this QRP by isolating the 99 projects where sample size did not change from dissertation to article. This is because those studies that added or deleted data have a logical (but not necessarily appropriate) reason why descriptive statistics would change. There were also 22 cases where not enough information was provided in the dissertation or article to determine if data had been altered (i.e., either the dissertation or publication did not include descriptives or a correlations matrix). Of the 77 remaining pairs, 25 (32.5%) showed changes in the means, standard deviations, or interrelations of the included variables.

Among the studies where data changed, there were 47 unsupported dissertation hypotheses and 63 supported dissertation hypotheses. When published, 16 (34.0%) of the unsupported hypotheses became statistically significant and none (0.0%) of the 63 supported hypotheses became statistically nonsignificant. The post hoc altering of data was associated with statistically nonsignificant dissertation hypotheses becoming significant journal hypotheses but had no effect on supported dissertation hypotheses.

Selective deletion or addition of variables. The same 22 cases where we could not determine if data had been altered were also excluded here, as we also could not determine if variables had been added or deleted with the information provided. Of the remaining 120 dissertation-publication pairs, there were 90 (75.0%) instances where not all of the variables included in the dissertation appeared in the publication and 63 (52.5%) instances where not all of the variables found in the article were reported in the dissertation. The overlap between the two groups was extensive, with 59 projects that both added and dropped variables.

The dropping of variables sometimes occurs due to the natural culling of material from the dissertation to the manuscript submission and revision, but the dropping of variables should be unrelated to statistical significance. Although it would be preferable for authors to acknowledge that not all the variables collected were included in the finished product, this QRP does not contribute to the Chrysalis Effect unless it coincides with a changing of the ratio of statistically significant to statistically nonsignificant hypotheses.

Among the studies that added variables, there were 84 unsupported dissertation hypotheses and 136 supported dissertation hypotheses. After the variables were added, 25 (29.8%) of the unsupported hypotheses became statistically significant versus 11 (8.1%) of the supported hypotheses becoming statistically nonsignificant. As with the other QRPs, the addition of variables coincided with increases in the number of statistically significant hypotheses (RR = 3.68; 95% CI = [1.91, 7.08]).

The dropped variables were often control variables not included in formal hypotheses. As such, there were a substantial number of dissertation hypotheses (361) that were retained in the journal article (i.e., common hypotheses). One hundred and forty-five of these retained hypotheses were not statistically significant and 216 were. Upon publication, 36 (24.7%) of the statistically nonsignificant dissertation hypotheses became statistically significant, and 15 (7.0%) of the statistically significant dissertation hypotheses changed to statistically nonsignificant journal hypotheses. Dropping variables was more than three times as likely to result in unsupported hypotheses becoming supported than vice versa (RR = 3.53; 95% CI = [2.01, 6.22]).

Reversing the direction or reframing hypotheses to support data. This QRP was the least common. We documented only eight projects (22 hypotheses) that maintained the same constructs in the dissertation hypothesis but had fundamentally changed the nature of the hypothesis in the journal article by altering the direction of the proposed effect or adding a boundary condition (e.g., "*X* relates to *Y*" in the dissertation changes into "*X* relates to *Y* when *Z* is controlled for" in the journal article). None of the 22 altered hypotheses were statistically significant in the dissertation, but upon publication, 17 (77.3%) changed from unsupported to supported. Ten of these instances were published in one of our leading outlets, *Academy of Management Journal*.

QRPs Associated With Unique (Dropped or Added) Hypotheses

The above QRPs were documented by observing the common hypotheses. However, some of the largest differences were among the hypotheses that appeared only in the dissertation or only in the journal article. One hundred and twenty-six (88.7%) projects either dropped or added hypotheses with the majority (80) doing both. If the adding and dropping of hypotheses were unrelated to statistical significance, then the ratio of supported to

unsupported should be approximately the same for dropped hypotheses as it is for added hypotheses. As shown in Table 2, there were 1,333 dropped hypotheses and 516 (38.7%) were statistically significant. This is a ratio of 0.63 to 1.00, which indicates that dropped hypotheses are a little more than 1.5 times as likely to be statistically nonsignificant as statistically significant.

The results follow an opposite pattern for the hypotheses unique to journal publications (i.e., added hypotheses). There were 333 hypotheses that were added after the dissertation was completed, and 233 (70.0%) were statistically significant. This corresponds to a ratio of supported to unsupported hypotheses of 2.33 to 1.00. Meaning that for every statistically nonsignificant hypothesis that was added, there were 2.33 added hypotheses that were statistically significant. Added hypotheses were nearly twice as likely to be statistically significant as dropped hypotheses (70.0% versus 38.7%). As with the QRPs tested within the retained or common hypotheses, the adding and dropping of hypotheses coincided with marked increases in the ratio of statistically significant to nonsignificant results (*RR* = 1.81; 95% CI = [1.64, 1.99]).

Overall Chrysalis Effect

Of the 1,978 hypotheses contained in the dissertations (i.e., dropped and common hypotheses), 889 (44.9%) were statistically significant.[1] That less than half of the hypotheses contained in a dissertation are supported with statistical significance is troubling, but more troubling is that 645 of the 978 (65.9%) hypotheses published in the journal articles (i.e., added and common hypotheses) were statistically significant. This is a 21.0% inflation of statistically significant results and corresponds to more than a doubling of the ratio of supported to unsupported hypotheses from 0.82:1 in the dissertations to 1.94:1 in the journal articles. To our knowledge, this is the first direct documentation of the prevalence, severity, and effect of QRPs in management research, and on the basis of these findings, we conclude that the published literature, at least as it relates to those early research efforts by junior faculty, is overstating its predictive accuracy by a substantial margin. Thus, we find evidence to support a Chrysalis Effect in management and applied psychology.

Antecedents and Outcomes of QRPs

In Table 3, we consider potential antecedents (e.g., initial ratio of supported to unsupported hypotheses) and outcomes (e.g., change in the ratio of supported to unsupported hypotheses; quality of the journal outlet where the dissertation data were published) of engagement in QRPs. We described eight (8) QRPs in our results (i.e., adding data, dropping data, altering data, adding variables, dropping variables, adding hypotheses, dropping hypotheses, and changing hypotheses) and identified another in our methods (i.e., failure to acknowledge the dissertation in the journal article). Since we could not detect cases where subjects were both added and dropped in a single project, the maximum number of QRPs we could identify for any one project was eight (8). There were zero (0) instances of individuals engaging in no QRPs and there were two (2) instances of individuals engaging in the maximum number of QRPs. The mode was 3 (n = 26) and the mean was 4.07 (SD = 1.89).

Table 3
Means, Standard Deviations, and Intercorrelations of QRP Engagement at Project Level

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. QRPs	4.07	1.89										
2. Committee as coauthor[a]	0.72	0.45	.13									
3. Dissertation date	2006	2.76	–.21*	.09								
4. Publication date	2009	2.39	–.06	.14	.74**							
5. Time to publication	3.39	1.89	.23**	.05	–.53**	.19*						
6. Affiliation at dissertation[b]	0.58	0.49	.13	–.02	–.19*	–.15	.09					
7. Affiliation at publication[b]	0.25	0.43	–.04	.03	.11	.04	–.11	.12				
8. Journal impact	50.18	42.96	.25**	–.04	–.13	–.17*	–.03	.15	.17*			
9. Dissertation ratio	0.51	0.30	–.22**	.11	.12	.02	–.15	–.19*	.06	.01		
10. Publication ratio	0.68	0.31	.08	.06	–.08	–.12	–.04	–.13	.07	.17*	.49**	
11. ΔD-P ratio	0.17	0.31	.28**	–.04	–.20*	–.13	.12	.07	.01	.16	–.48**	.53**

Note: n = 142. QRP = questionable research practice; ratio = ratio of supported to unsupported hypotheses; ΔD-P ratio = change in the ratio of supported to unsupported hypotheses from dissertation to journal publication.
[a]Committee as coauthor coded 1 if a committee member was a coauthor and 0 otherwise.
[b]Institutions rated as "very high research activity" by the Carnegie Classification of Institutions of Higher Education were coded as 1; all others were coded as 0.
*p < .05.
**p < .01.

As shown in Table 3, QRPs were more prevalent when the ratio of supported to unsupported hypotheses in the dissertation was low (r = –.21). In addition, QRP engagement increased the longer the dissertation remained unpublished (r = .23). These findings are consistent with general strain theory as "poor" results create goal blockage in the form of inability to publish one's dissertation, which leads to goal pursuit through non-ideal channels.

This begs the questions of whether this pursuit does in fact improve the ratio of supported to unsupported hypotheses and whether these improved findings actually lead to goal attainment. Based on these data, the answer to both questions is yes. QRPs were associated with improvements to the ratio of supported to unsupported hypotheses from dissertation to journal article (ΔD-P ratio, r = .28) as well as publication in more prestigious outlets (r = .25). Neither the research intensity of the doctoral granting institution (r = .13) nor the research intensity of the author's affiliation at time of publication (r = –.04) had much effect at mitigating QRPs. The inclusion of a dissertation committee member on the publication also did little to reduce QRPs, and the correlation was in fact positive (r = .13). The one hopeful finding was that more recent projects contained fewer QRPs (r = –.21). This may indicate less engagement in QRPs among recent graduates, but it is also possible that those dissertations without a high ratio of statistically significant to nonsignificant hypotheses have not yet found their way into the literature via QRPs.

Discussion

Using general strain theory as an explanatory framework, we posited that the current reward system and absence of checks and balances in the publication process create a strong

motive and opportunity for researchers to engage in QRPs as a means to better their chances of publication. Publications are the primary method of goal attainment (e.g., tenure, reputation), and when hypotheses are not supported, this goal is, or at least is perceived to be, blocked. Goal blockage creates strain that can motivate a researcher to achieve his or her goal through inappropriate channels, including QRPs. The collective result of QRPs is a biasing of the literature that we label the Chrysalis Effect. We tested and found support for the Chrysalis Effect by tracking the path to publication of 142 dissertations. We also found small to moderate relations between engagement in QRPs and journal quality, time to publication, "poor" initial results, and improvements to these initial results in terms of statistical significance.

Theoretical and Practical Implications

The chief implication of the Chrysalis Effect is the uncertainty with which we can view the management literature. The more than 20% jump in statistically significant hypotheses upwardly biases the validity of management theories tested in our sample by systematically excluding cases where theories and hypotheses were unsupported by statistical significance tests. Because of the strong reliance on past research to guide theory development (Hambrick, 2007), an inaccurate literature hampers the ability to generate and test new theory or to make incremental contributions to existing theory. Further, it is not a consistent upward bias, and the variance in QRP engagement compromises the ability to distinguish theory supported with best practices in hypothesis development, data collection, analysis, and presentation from theory supported with QRPs. Perhaps more troubling is that the prevalence of QRPs suggests that they have been institutionalized. In recent years, many industries where nonideal practices were widespread and systematic have experienced significant challenges and opened themselves to external oversight. For example, Major League Baseball (MLB) experienced a surge in home runs hit by players during the 1990s and early 2000s. Although initially turning a blind eye to the possibility of steroids accounting for the increase in ability, whistleblowers and government intervention has since caused MLB to take a hard stance against steroids and enforce stringent drug-testing policies with accompanying sanctions (Banks & O'Boyle, 2013). If we cannot self-police by establishing and enforcing best practices, then those external stakeholders that provide funding (e.g., state governments, federal grant agencies, private individuals and organizations) may reduce or withdraw their support.

In addition, the Chrysalis Effect likely results in continued construct and theory proliferation (Harter & Schmidt, 2008). Again, management research rarely engages in replication studies. Consequently, few engage in theory pruning and the result of this is an unwieldy literature that creates far more constructs and theory than it prunes (Leavitt, Mitchell, & Peterson, 2010). QRPs help statistically nonsignificant results become significant and lead to nearly all theory being supported, all measures being validated, and all interventions being successful. As such, management, like other sciences, may be experiencing a suboptimal rise in the number of supported hypotheses as unsupported hypotheses all but disappear from the literature (Fanelli, 2012).

Our findings also contribute to the debate over the trustworthiness of the management literature due to issues such as publication and outcome-reporting bias, and we diverge from some calls that the published literature is an entirely accurate representation of the population of research. For example, Dalton et al. (2012) found "that, contrary to the established belief,

the file drawer problem is of little, if any, consequence for meta-analytically derived theoretical conclusions and applications in OBHRM, I-O psychology, and related fields" (p. 225). Their technique compared the mean effect size of the correlation matrices in published studies to the mean effect size of the correlation matrices in unpublished studies. However, as illustrated in this study, QRPs more likely target certain relations and variables, and engagement in QRPs is unlikely an attempt to raise the mean value of all effect sizes in a study. We expect that the file drawer problem follows a similar pattern where studies are suppressed only if and when key relations fail to achieve statistical significance. Our belief, and an area of future research, is that if one were to collect a representative sample of published and unpublished work and isolate those effect sizes that were explicitly hypothesized, one may arrive at a different conclusion than Dalton et al. In sum, we echo the call of many (e.g., Kepes & McDaniel, 2013) and strongly encourage greater investigation into both publication bias and outcome reporting bias.

Our findings also inform the possibility that not only is there a suboptimal rise in statistical significance, but there may also be a suboptimal rise in effect size magnitudes. Whereas adding, dropping, and altering hypotheses should not influence effect size magnitude, QRPs that alter data can inflate effect sizes and potentially bias systematic reviews. This is because dropping and altering data can increase the likelihood of statistical significance only by way of increasing effect sizes. All things equal, adding data should not increase effect sizes, but selectively adding data (e.g., collecting 20 more subjects and adding only those 15 that increase the effect size) can certainly upwardly bias systematic review estimates. Thus, any change to the data for the purpose of achieving statistical significance has the very real potential to inflate meta-analytically derived effect size estimates.

Engagement in QRPs also has negative consequences for practice. Medical researchers have already recognized the important implications of QRPs for practitioners and have engaged in studies that compared databases, registries, and protocols against published journal articles for evidence of engagement in QRPs (e.g., Chan, Hrobjartsson, Jorgensen, Gotzsche, & Altman, 2008; Dwan et al., 2008, 2011; Huic, Marusic, & Marusic, 2011; Mathieu, Boutron, Moher, Altman, & Ravaud, 2009; Turner, Knoepflmacher, & Shapley, 2012). While QRPs in management will typically not have the same immediate negative consequences for human health and well-being, the implications can still be harmful. For example, one dissertation in this study considered the relationship between high-performance work systems and firm-level outcomes. Another investigated the various means to reduce supervisor aggression. The findings of these studies have important implications for firm financial performance and employee well-being. Practitioners may make decisions based upon the incomplete evidence reported in the journal article. Consequently, engagement in QRPs misguides practitioners.

Encouraging evidence-based practices that are founded on questionable or incorrect evidence is likely to hurt the credibility of the field for both research and practice. In the case of the former, there already exist well-documented perceptions that the social sciences are as a whole inferior to "harder" areas of science (Fanelli, 2010b). Widespread engagement in QRPs reinforces this stereotype. In terms of the latter, QRPs are likely to widen the often lamented gap between science and practice (Banks & McDaniel, 2011; Briner & Rousseau, 2011). Furthermore, as engagement in QRPs can lead to publication bias and outcome-reporting bias (Chan et al., 2008; Dwan et al., 2008), research has provided evidence that

engagement in QRPs may also bias meta-analytic estimates (Hahn, Williamson, Hutton, Garner, & Flynn, 2000). As meta-analytic results provide important guidance for evidence-based practice (Briner & Rousseau, 2011; Le, Oh, Shaffer, & Schmidt, 2007), more research that considers how to deal with outcome-reporting bias within the context of meta-analysis is needed (e.g., Hahn et al., 2000; Kirkham, Riley, & Williams, 2011). We suggest that unless legitimate efforts are made to reduce the prevalence of engagement in QRPs, the credibility of management and the social sciences in general will fail to rise in the hierarchy of science and evidence-based practice.

We do wish to clarify that the QRPs identified here were largely questionable because they were not reported or were incorrectly reported. If post hoc changes had been reported as post hoc, then their appropriateness would be under the purview of the reader. It is the lack of information and in some cases, such as presenting post hoc findings as a priori hypotheses, misinformation that data mining or "peeking" becomes a QRP. Further, *questionable* is not equivalent to *inappropriate*. Between the dissertation defense and journal publication, it is possible, even likely, that mistakes were identified and corrected, outliers removed, new analytic techniques employed, and so on that would be classified as questionable by our criteria but were nevertheless wholly appropriate to that particular project. That being said, these changes consistently coincided with increases in statistical significance and increases in the ratio of supported to unsupported hypotheses, and on this basis, we conclude that the preponderance of QRPs are engaged in for non-ideal reasons.

Suggestions for the Reduction of QRPs

Despite increased awareness of what QRPs are and the damage they cause (e.g., Bedeian et al., 2010; Martinson, Anderson, & De Vries, 2005), QRPs persist. We contend that this is because as a field, we reward QRPs, and we are embedded within a culture that reduces the likelihood of their detection. As such, QRP reductions are unlikely to occur by placing the onus of best practices on the individual researcher. Asking researchers to forego immediate, extrinsic rewards in order to serve the higher ideals of fair play and professional ethics is a noble request but one that is unlikely to manifest into real change. That QRPs have the end result of a compromised science akin to the "tragedy of the commons" (Hardin, 1968) is likely insufficient motivation for reducing their prevalence. As long as researchers perceive goal blockage, they will experience strain that will generate negative emotions and encourage goal attainment through QRPs. Given that we are attempting to change the norms and culture of the field rather than directly influence the motivations and perceptions of its members, most of our recommendations require support from established academics who hold leadership positions in journals and professional associations. These individuals are the first movers, and we humbly suggest that without their action, the Chrysalis Effect will persist.

Our first suggestion for institutional leaders is by far the quickest and easiest to implement. Upon manuscript submission, researchers affirm that they did not engage in any of the specific QRPs discussed here (e.g., dropping data) or elsewhere (e.g., "cherry-picking" fit indices). An abbreviated form is already present at some outlets, such as *Journal of Applied Psychology*. However, it would be helpful if the process were standardized, required personal signatures, and applied to all authors on the submission, not just the corresponding author.

The implementation of an "honor code" or "commitment reminders" would reduce QRPs in two ways. First, it prevents engagement in QRPs out of ignorance. If a researcher were

unaware that a certain practice was questionable, he or she would be informed in the submission and could make the necessary changes or acknowledgments. Empirical evidence has shown that decreasing the malleability to interpret one's behavior results in reduced acts of dishonesty (Mazar, Amir, & Ariely, 2008). Second, an author's affirmation that QRPs did not occur constitutes a statement of compliance. The act of acknowledging conformance to proscribed norms reduces some forms of inappropriate behavior, such as academic dishonesty (Ariely, 2012; Fischer, 1970; McCabe, Trevino, & Butterfield, 1996). Further, codes of conduct lend themselves to practices more consistent with scientific ideals (Kish-Gephart, Harrison, & Treviño, 2010; Stroebe, Postmes, & Spears, 2012), and if violated, the journal would have the ability to ask that the manuscript be withdrawn or publication be corrected with an erratum. It seems appropriate that as gatekeepers of scientific knowledge, journals should lead this charge.

Our second suggestion, and one consistent with recommendations by Tsang and Frey (2007), is that the original submission of any published article be available for download. One limitation to our research is that we possessed only the completed dissertation and the final version of the journal article. Thus, we could not establish at what point changes occurred, which means we could not determine whether the author made the changes preemptively because of perceived reviewer impressions or instead as a concession to actual reviewer comments. Although we suspect that most changes occur preemptively, if the original submission can be compared to the finished product, then there is accountability on the reviewers and action editor to suggest only changes that are consistent with best practices. We still contend that peer review is a very positive process, but it is always a post hoc process, and changes that occur throughout the revision should be documented.

Our third suggestion is that some journal space in top-tier outlets be specifically reserved for replications. Although many top journals express their willingness to publish replications (e.g., Eden, 2002), little of this work is reflected in these journals. This would indicate that replications either are not submitted or are not favorably reviewed. As such, there is a need to encourage replication studies, and we suggest a fixed schedule of reinforcement (i.e., guaranteed journal space in every issue). We appreciate that journal space is already limited and many of our top journals are able to fit only seven or so articles in an issue. However, replications need not include an extensive literature review, hypothesis development, and so on. A summary of the results in the original manuscript and a summary of the results of the replication could take up as little as two to three pages. Initial steps are already underway with the creation of *Academy of Management Discoveries*, which encourages triangulations and expanded replications of research findings to establish boundary conditions.[2] However, more outlets, especially established outlets (e.g., *Journal of Management, Journal of Applied Psychology, Strategic Management Journal*) should also encourage replications. This would motivate researchers to replicate findings as their work has the potential to appear in top-tier journals. Although a vita that contained only replication studies would likely not result in a positive tenure decision at most schools, replication studies when coupled with an active research stream of original work could bolster one's chances. An additional advantage of space allocation for replication studies is that it has the potential to spark debate and help to winnow unnecessary theory and constructs.

If the third suggestion was the carrot, then the fourth suggestion is the stick. Editors need to establish and enforce data-sharing policies. Good science possesses the property of falsification (Popper, 1968). Replication with new data serves as one form of falsification, but a

second form of falsification is replication with the original data. If a study, due to incomplete reporting or an unwillingness to provide the original data, cannot be replicated, then it lacks an important property of good science. The National Academy of Sciences and the National Institute of Health have taken up this cause, and their policies require that biomedical researchers receiving funding share their data after a 12-month period of exclusivity (Guttmacher, Nabel, & Collins, 2009).

There are few instances in management where data are so sensitive that they cannot be provided to another researcher for the purposes of replication. Even data collected from private companies or from the military can usually be stripped of identifiers so that anonymity of participants is preserved. If data are so sensitive or the organization(s) that provide the data so stringent about proprietary rights, then perhaps in the interest of confidentiality, findings derived from those data should not be published in a journal. At the very least, the inability to provide the data for replication should be acknowledged in the manuscript submission, and it should factor into the editor's decision to accept or reject the paper. We also suggest a journal policy whereby if authors do not acknowledge the data privacy issues and still refuse or are otherwise unable to provide the original data for a given time window, then the publication should be retracted.

Summary of recommendations. Some of these suggestions are more difficult to implement than others, but we have tried to focus on those institutional changes that are reasonable and possible in the near future. Yet, we do not wish to discount those long-term, paradigm-changing solutions, such as research registries (Banks & McDaniel, 2011), blinding reviewers to the results and discussion sections (Kepes & McDaniel, 2013), deemphasizing the sometimes rigid adherence to theory-driven hypotheses (Hambrick, 2007), and deinstitutionalizing null hypothesis significance testing (Orlitzky, 2012; Schmidt, 1996). These seismic shifts may come about eventually, but in the meantime, we believe these incremental changes will increase the quality and veracity of the management literature.

Limitations and Future Direction

One important limitation to this work is that it focused on the means of goal attainment within the hypothetico-deductive domain. Because the extant literature is primarily composed of quantitative research with null hypothesis significance testing, QRPs in this domain are likely the most common. However, QRPs can be found across all methodologies and analytic techniques. For example, in structural equation modeling, overall fit is one determinant of theory adequacy. Rather than being concerned about rejecting a null hypothesis, a researcher may be tempted to engage in QRPs that improve model fit to acceptable thresholds. QRPs in this domain would include "cherry-picking" fit indices, data-driven alterations to models, and failing to report estimated paths (e.g., correlated residuals). Likewise, Bayesian statistics allow for the specification of an a priori distribution, which can encourage researchers to test a variety of distributions and choose the one that best fits their preconceived notions. Thus, our results speak only to the QRPs associated with increasing the ratio of statistically significant to nonsignificant hypotheses, and future research should consider other domains.

There is another important limitation regarding the generalizability of our results. Our data were confined to junior faculty and doctoral students, and more specifically, to those

attempting to publish their dissertations. Thus, a substantial amount of the population of research is systematically excluded from this work, and we discourage generalizing any specific finding to the entirety of the field (e.g., all management research is upwardly biased by 21%). It is possible that engagement in QRPs as a dissertation becomes a journal publication is systematically higher or lower than what occurs in the population of research. The QRPs we document here may not persist to other research projects, especially once the researcher arrives at a point where strain may be reduced (e.g., tenure). It is also possible that the reward system of academia encourages QRPs only by doctoral students and junior faculty. Yet, it may also be the case that during these formative years, academics develop the habits that guide their research, mentoring, and reviewing for the rest of their career. Perhaps even after tenure, there are enough goals perceived to be blocked by statistically nonsignificant results (e.g., reputation, job mobility) that the strain is still strong enough to engage in QRPs. These are future directions of research that our work cannot definitively answer. However, our work clearly indicates these are important and worthwhile avenues to pursue.

A third limitation of the current work is that despite finding evidence of the Chrysalis Effect, the tests for it were not very sensitive, nor do they offer insight into the actual process by which the practices emerged. If one considers the full timeline of a research project from initial idea to publication, the defended dissertation is very far along in the process, and any QRPs that occurred prior to dissertation defense could not be documented. Relatedly, the gap between defense and publication contains a manuscript submission stage(s) and often multiple rounds of reviews. Decisions to engage in QRPs could have occurred any time during this process, and it is not clear whether institutional pressure, editorial pressure, or self-inflicted pressure motivated the QRPs. Although we found a substantially higher incidence of supported hypotheses in the articles than the dissertation, the problem of a Chrysalis Effect may be even worse due to the QRPs we were unable to document. Further, certain discrepancies between dissertations and subsequent publications were treated as QRPs (e.g., dropping hypotheses) when these behaviors might be necessary to conform to journal space or to refine a theory. Therefore, although we found evidence for our hypothesized effect by the change in the ratio of supported to unsupported hypotheses, we are unable to determine the cause of the discrepancies. Future research should seek out new methods and techniques to determine the full extent of the Chrysalis Effect. For example, future research could use in-depth case analysis to explore the decision-making process as it unfolds, perhaps with an agreement of anonymity among a small group of scholars interviewed at multiple stages in their dissertation and subsequent publication efforts.

Finally, our research was limited by its sample size in that we did not have the ability to examine subfields and specific content areas of management. The Chrysalis Effect may be stronger in nascent fields, such as entrepreneurship, that have yet to benefit from decades of theoretical grounding (Crook, Shook, Morris, & Madden, 2010). Alternatively, the Chrysalis Effect may be stronger in disciplines such as organizational behavior and human resources because established fields have a greater theoretical base to draw from to justify hypothesis reversal and other post hoc alterations.

Conclusion

We reviewed the means, motives, and opportunity to improve the likelihood of publication independent of rigor and relevance. We then documented QRP incidence by tracking

changes between dissertations and resulting publications and found a marked increase in statistical significance from dissertation to journal article. Our conclusion is that the published literature overestimates the predictive accuracy of management science and reflects a form of publication bias referred to as the Chrysalis Effect.

Notes

1. To assess the representativeness of the dissertations, we drew a sample of 20 management dissertations with no resulting publications. Of the 228 hypotheses included in these dissertations, 109 (47.9%) were supported with statistical significance.

2. Triangulation is characterized as the use of "multiple reference points to locate an object's exact position" (Jick, 1979, p. 602). In the context of management science, triangulation is the use of multiple study designs, settings, samples, and methods to investigate a particular research question (P. Sackett & Larson, 1990).

References

Agnew, R. 1992. Foundation for a general strain theory of crime and delinquency. *Criminology*, 30: 47-88.

Agnew, R., Brezina, T., Wright, J. P., & Cullen, F. T. 2002. Strain, personality traits, and delinquency: Extending general strain theory. *Criminology*, 40: 43-72.

Aguinis, H., Gottfredson, R. K., & Joo, H. in press. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*.

Academy of Management. 2006. Academy of Management code of ethics. Retrieved March 1, 2013, from http://aom.org/uploadedFiles/About_AOM/Governance/AOM_Code_of_Ethics.pdf

American Psychological Association. 2010. American Psychological Association ethical principles of psychologists and code of conduct. Retrieved March 1, 2013, from http://www.apa.org/ethics/code2002.html

Ariely, D. 2012. *The (honest) truth about dishonesty: How we lie to everyone-especially ourselves*. New York, NY: HarperCollins.

Bakker, M., van Dijk, A., & Wicherts, J. M. 2012. The rules of the game called psychological science. *Perspectives on Psychological Science*, 7: 543-554.

Banks, G. C., & McDaniel, M. A. 2011. The kryptonite of evidence-based I-O psychology. *Industrial and Organizational Psychology*, 4: 40-44.

Banks, G. C., & O'Boyle, E. H. 2013. Why we need I-O psychology to fix I-O psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 6: 284-287.

Bedeian, A. G. 2003. The manuscript review process: The proper roles of authors, referees, and editors. *Journal of Management Inquiry*, 12: 331-338.

Bedeian, A. G. 2004. Peer review and the social construction of knowledge in the management discipline. *Academy of Management Learning & Education*, 3: 198-216.

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.

Bracken, M. B., & Sinclair, J. C. (1998). When can odds ratios mislead? Avoidable systematic error in estimating treatment effects must not be tolerated. *British Medical Journal*, 317:1156-1157.

Briner, R. B., & Rousseau, D. M. 2011. Evidence-based I-O psychology: Not there yet. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 4: 3-22.

Buchheit, S., Collins, D., & Reitenga, A. 2002. A cross-discipline comparison of top-tier academic journal publication rates: 1997-1999. *Journal of Accounting Education*, 20: 123-130.

Cashen, L. H., & Geiger, S. W. 2004. Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods*, 7: 151-167.

Chalmers, I. 1990. Underreporting research is scientific misconduct. *Journal of the American Medical Association*, 263: 1405-1408.

Chan, A. W., Hrobjartsson, A., Jorgensen, K. J., Gotzsche, P. C., & Altman, D. G. 2008. Discrepancies in sample size calculations and data analyses reported in randomised trials: Comparison of publications with protocols. *British Medical Journal*, 337, 1-8.

Chen, X. P. 2011. Author ethical dilemmas in the research publication process. *Management and Organization Review*, 7: 423-432.

Cloward, R. A., & Ohlin, L. E. 1960. *Delinquency and opportunity*. New York, NY: Free Press.

Cohen, A. K. 1955. *Delinquent boys*. New York, NY: Free Press.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37-46.

Cohen, J. 1990. Things I have learned (thus far). *American Psychologist*, 45: 1304-1312.

Cohen, J. 1994. The earth is round ($p < .05$). *American Psychologist*, 49: 997-1003.

Corley, K. G., & Gioia, D. A. 2000. The rankings game: Managing business school reputation. *Corporate Reputation Review*, 3: 319-333.

Crook, T. R., Shook, C. L., Morris, M. L., & Madden, T. M. 2010. Are we there yet? An assessment of research design and construct measurement practices in entrepreneurship research. *Organizational Research Methods*, 13: 192-206.

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. 2012. Revisiting the file drawer problem in meta-analysis: An assessment of published and nonpublished correlation matrices. *Personnel Psychology*, 65: 221-249.

Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? *British Medical Journal*, 316: 989-991.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A. W., Cronin, E., & Williamson, P. R. 2008. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*, 3: e3081.

Dwan, K., Altman, D. G., Cresswell, L., Blundell, M., Gamble, C., & Williams, P. R. 2011. Comparison of protocols and registry entries to published reports for randomised controlled trials. The Cochrane Collaboration.

Eden, D. 2002. From the editors: Replication, meta-analysis, scientific progress, and *AMJ*'s publication policy. *Academy of Management Journal*, 45: 841-846.

Emerson, G. B., Warme, W. J., Wolf, F. M., Heckman, J., Brand, R. A., & Leopold, S. S. 2008. Testing for the presence of positive-outcome bias in peer review: A randomized controlled trial. *Archives of Internal Medicine*, 170: 1934-1939.

Fanelli, D. 2010a. Do pressures to publish increase scientists' bias? An empirical support from United States data. *PLoS One*, 5: e10271.

Fanelli, D. 2010b. "Positive" results increase down the hierarchy of the sciences. *PLoS One*, 5: e10068.

Fanelli, D. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90: 891-904.

Fischer, C. T. 1970. Levels of cheating under conditions of informative appeal to honesty, public affirmation of value, and threats of punishment. *Journal of Educational Research*, 64: 12-16.

Frank, M. C., & Saxe, R. 2012. Teaching replication. *Perspectives on Psychological Science*, 7: 600-604.

Gerber, A. S., & Malhotra, N. 2008. Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods and Research*, 37: 3-30.

Gigerenzer, G. 2004. Mindless statistics. *Journal of Socio-Economics*, 33: 587-606.

Gioia, D., & Corley, K. G. 2002. Being good versus looking good: Business school rankings and the Circean transformation from substance to image. *Academy of Management Learning and Education*, 1: 107-120.

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

Guttmacher, A. E., Nabel, E. G., & Collins, F. S. 2009. Why data-sharing policies matter. *Proceedings of the National Academy of Sciences*, 106: 16894.

Hahn, S., Williamson, P. R., Hutton, J. L., Garner, P., & Flynn, E. V. 2000. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Statistics in Medicine*, 19: 3325-3336.

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

Hardin, G. 1968. The tragedy of the commons. *Science*, 162, 1243-1248.

Harter, J. K., & Schmidt, F. L. 2008. Conceptual versus empirical distinctions among constructs: Implications for discriminant validity. *Industrial and Organizational Psychology*, 1: 36-39.

Harzing, A. W., & van der Wal, R. 2008. A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business. *Journal of the American Society for Information Science and Technology*, 60: 41-46.

Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102: 16569-16572.

Huic, M., Marusic, M., & Marusic, A. 2011. Completeness and changes in registered data and reporting bias of randomized controlled trials in ICMJE journals after trial registration policy. *PLoS One*, 6: 9.

Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLoS Medicine*, 2: e124.

Ioannidis, J. P. A. 2008. Why most discovered true associations are inflated. *Epidemiology*, 19: 640-648.

Jick, T. D. 1979. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 24: 602-611.

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.

Kacmar, K. M. 2009. From the editors: An ethical quiz. *Academy of Management Journal*, 52: 432-434.

Kemper, T. D. 1978. *A social interactional theory of emotions*. New York, NY: Wiley.

Kepes, S., & McDaniel, M. A. 2013. How trustworthy is the scientific literature in I-O psychology? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 6: 252-268.

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

Kirkham, J. J., Riley, R. D., & Williams, P. R. 2011. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in Medicine*, 31: 2179-2195.

Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. 2010. Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *Journal of Applied Psychology*, 95: 1-31.

Klein, O., Doyen, S., Leys, C., Miller, S., Questienne, L., & Cleeremans, A. 2012. Low hopes, high expectations, expectancy effects, and the replicability of behavioral experiments. *Perspectives on Psychological Science*, 7: 572-584.

Knol, M. J., Duijnhoven, R. G., Grobbee, D. E., Moons, K. G., & Groenwold, R. H. (2011). Potential misinterpretation of treatment effects due to use of odds ratios and logistic regression in randomized controlled trials. *PLoS One*, 6: e21248.

Leavitt, K., Mitchell, T. R., & Peterson, J. 2010. Theory pruning: Strategies to reduce our dense theoretical landscape. *Organizational Research Methods*, 13: 644-667.

Le, H., Oh, I.-S., Shaffer, J., & Schmidt, F. L. 2007. Implications of methodological advances for the practice of personnel selection: How practitioners benefit from meta-analysis. *Academy of Management Perspectives*, 21: 6-15.

Leung, K. 2011. Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. *Management and Organization Review*, 7: 85-94.

Locke, E. A., & Latham, G. P. 1990. *A theory of goal setting and task performance*. Upper Saddle River, NJ: Prentice Hall.

MacDonald, S., & Kam, J. 2007. Ring a ring o' roses: Quality journals and gamesmanship in management studies. *Journal of Management Studies*, 44: 640-55.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. 2011. Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, 112: 331-348.

Martinson, B. C., Anderson, M. S., & De Vries, R. 2005. Scientists behaving badly. *Nature*, 435: 737-738.

Masicampo, E. J., & Lalande, D. R. 2012. A peculiar prevalence of *p* values just below. 05. *Quarterly Journal of Experimental Psychology*, 65: 2271-2279.

Mathieu, S., Boutron, I., Moher, D., Altman, D. G., & Ravaud, P. 2009. Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA: The Journal of the American Medical Association*, 302: 977-984.

Mazar, N., Amir, O., & Ariely, D. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45: 633-644.

McCabe, D. L., Trevino, L. K., & Butterfield, K. D. 1996. The influence of collegiate and corporate codes of conduct on ethics-related behavior in the workplace. *Business Ethics Quarterly*, 6: 461-476.

Meehl, P. E. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46: 806-834.

Merton, R. K. 1938. Social structure and anomie. *American Sociological Review*, 3: 672-682.

Mills, J. L. 1993. Data torturing. *New England Journal of Medicine*, 329: 1196-1199.

Nicholas, J. M., & Katz, M. 1985. Research methods and reporting practices in organization development: A review and some guidelines. *Academy of Management Review*, 10: 737-749.

Nickerson, R. S. 2000. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5: 241-301.

Orlitzky, M. 2012. How can significance tests be deinstitutionalized? *Organizational Research Methods*, 15: 199-228.

Popper, K. R. 1968. *The logic of scientific discovery*. New York, NY: Harper and Row.

Rosenthal, R. 1990. Replication in behavioral research. *Journal of Social Behavior and Personality*, 5: 1-30.

Rynes, S., & Gerhart, B. 2003. *Compensation: Theory, evidence, and strategic implications*. Thousand Oaks, CA: Sage.

Sackett, D. L., Deeks, J. J., & Altman, D. G. (1996). Down with odds ratios! *Evidence Based Medicine*, 1: 164-166.

Sackett, P. R., & Larson, J. R. 1990. Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology*, vol. 1: 419-489. Palo Alto, CA: Consulting Psychologists Press.

Schmidt, F., & Hunter, J. 2002. Are there benefits from NHST? *American Psychologist*, 57: 65-66.

Schmidt, F. L. 1992. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, 47: 1173-1181.

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

Schminke, M. 2009. Editor's comments: The better angels of our nature-ethics and integrity in the publishing process. *Academy of Management Review*, 34: 586-591.

Siemens, J. C., Burton, S., Jensen, T., & Mendoza, N. A. T. 2005. An examination of the relationship between research productivity in prestigious business journals and popular press. *Journal of Business Research*, 58: 467-76.

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.

Stroebe, W., Postmes, T., & Spears, R. 2012. Scientific misconduct and the myth of self-correction in science. *Perspective of Psychological Science*, 7: 670-688.

Tsang, E. W., & Frey, B. S. 2007. The as-is journal review process: Let authors own their ideas. *Academy of Management Learning and Education*, 6: 128-136.

Turner, E. H., Knoepflmacher, D., & Shapley, L. 2012. Publication bias in antipsychotic trials: An analysis of efficacy comparing the published literature to the US Food and Drug Administration database. *PLoS Medicine*, 9: 1-17.

Wagenmakers, E. J. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, 14: 779-804.

Wood, J. A. 2008. Methodology for dealing with duplicate study effects in a meta-analysis. *Organizational Research Methods*, 11: 79-95.

Journal of Management
Vol. XX No. X, Month XXXX 1
DOI: 10.1177/0149206316667701

Corrigendum

O'Boyle, E. H., Jr., Banks, G. C., & Gonzalez-Mulé, E. In press. The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*. [Epub ahead of print March 19, 2014.] (Original DOI: 10.1177/0149206314527133)

In Table 1 in the initial OnlineFirst version of this article, the reported risk ratio (RR) was incorrectly listed as 2.35 in the "Add/drop" data row. The correct value is 2.41, and the associated 95% confidence interval [1.16, 5.01] and statistical significance (p = .02) remain unchanged. Table 1 and the last sentence of the first full paragraph on page 11 are updated with the correct value in the latest online version.