Strategic Management Journal
Strat. Mgmt. J., **37**: 167–176 (2016)
Published online EarlyView in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/smj.2459
Received 31 August 2013; *Final revision received 26 February 2015*

SCIENTIFIC APOPHENIA IN STRATEGIC MANAGEMENT RESEARCH: SIGNIFICANCE TESTS & MISTAKEN INFERENCE

BRENT GOLDFARB[1]* and ANDREW A. KING[2]
[1] *Department of Management and Organizations, Robert H. Smith School of Business, University of Maryland, College Park, Maryland, U.S.A.*
[2] *Tuck School of Business at Dartmouth, Dartmouth College, Tuck School of Business, Hanover, New Hampshire, U.S.A.*

Research summary: *This article uses distributional matching and posterior predictive checks to estimate the extent of false and inflated findings in empirical research on strategic management. Based on a sample of 300 papers in top outlets for research on strategic management, we estimate that if each study were repeated, 24–40 percent of significant coefficients would become insignificant at the five percent level. Our best guess is that for about half of these, the true coefficient is very close to 0. The remaining coefficients are likely directionally correct but inflated in magnitude. We offer several practical individual and field level suggestions for reducing* scientific apophenia, *that is, our tendency to find and publish evidence of order where none exists.*

Managerial summary: *This article analyzes the degree to which statistics in research on strategic management provide meaningful evidence for decision-making. Based on a sample of 300 papers, we estimate that 24%–40% of reported results would probably not be confirmed if the study were repeated. Our best guess is that about half of the reported results are wrong ($B\sim=0$) and the other half of results too weak to find repeatedly. We conclude that scientific apophenia—the tendency to find evidence of order where none exists—is a serious problem in the literature on strategic management. We recommend replication of empirical studies to insure that they provide evidence for guiding managers. We also provide guidance for avoiding scientific apophenia in empirical research.*

> Human understanding is of its own nature prone to suppose the existence of more order and regularity in the world than it finds. —(Francis Bacon, (1620) Novum Organum, Aphorism XLV)

INTRODUCTION

In a recent issue of this journal, Bettis (2012) reports a conversation with a graduate student who forthrightly announced that he had been trained by faculty to "search for asterisks." The student explained that he sifted through large databases for statistically significant results and "[w]hen such models were found, he helped his mentors propose theories and hypotheses on the basis of which the 'asterisks' might be explained" (p. 109). Such an approach, Bettis notes, is an excellent way to find seemingly meaningful patterns in random data. He expresses concern that these practices are common, but notes that unfortunately "we simply do not have any baseline data on how big or small are the problems" (Bettis, 2012: 112).

In this article, we address the need for empirical evidence by exploring the extent of what we term "scientific apophenia" in research on strategic

Keywords: empirical analysis; statistical inference; false positives; specification searches; confirmatory research
*Correspondence to: Brent Goldfarb, 7621 Mowatt Ln, College Park, MD 20740. E-mail: bgoldfarb@rhsmith.umd.edu

management. The term *apophenia* has been used in clinical psychology to mean the perception of "connection or meaningful pattern between unrelated or random things (Merriam-Webster, 2014)." In our context, we use it to define not a medical disorder, but a dysfunction in the way scientists find meaning in data. We define *scientific apophenia* as the assigning of inferential meaning when limited statistical power should have prevented such a conclusion or when the data are actually random. Our definition fits the "three-valued logic of Kaiser, Tukey, Abelson, Harris, Tyron and Cox" that suggests forming three results from tests (true, false, or unknown) (Hurlbert and Lombardi, 2009: 311).

Scholars propose that scientific apophenia results from three related processes: authors or reviewers search repeatedly for statistically significant results (Bettis, 2012; Denton, 1985); samples and models are manipulated to nudge results across significant thresholds (Simmons, Nelson, and Simonsohn, 2011); and readers make inferences from published estimates that are unreliable or inflated (Stanley, 2005).

Bettis (2012) reports that computer power now allows researchers to sift repeatedly through data in search of patterns. Such specification searches can greatly increase the probability of finding an apparently meaningful relationship in random data. For example, in regression analyses that we conducted on two randomly generated values (Y and X), we discovered that by trying four functional forms for X, a researcher can increase the chance of a false positive for a significant relationship between Y & X from 1 in 20 to about 1 in 6. STATA code for this simulation is available in Appendix S1. In other words, biasing selection on "significant" findings can occur at either the author or reviewer level. Authors can sift through data until they find and submit "significant results," or journal reviewers can strain through submissions and accept only those with "meaningful" findings. The effect is the same regardless of the location of the selection process (Denton, 1985).

Simmons *et al.* (2011) contend that some authors also push almost significant results over thresholds by removing observations or gathering more data, by dropping experimental conditions, by adding covariates to specified models, and so on. The physicist Richard Feynman claims such behavior explains why repeated experimental estimates of the charge of the electron were biased by the original, but erroneous, estimate made by Nobel Laureate Robert Milliken. "When [experimenters] got a number that was too high above Millikan's, they thought something must be wrong—and they would look for and find a reason why something might be wrong. When they got a number close to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off" (Feynman, 1985: 342).

If readers of published research fail to account for the above tendencies, then they will give undue credence to estimated coefficients. Doing so is a human tendency, one of our reviewers suggested, because when evidence is uncertain, a single example is often considered representative of the whole (Tversky and Kahneman, 1973). Such inference is incorrect, however, if selection occurs on significant results. In fact, if "significant" results are more likely to be published, coefficient estimates will inflate the true magnitude of the studied effect—particularly if a low powered test has been used (Stanley, 2005).

To what extent is scientific apophenia present in the literature on strategic management? In this article, we analyze a random sample of manuscripts from top outlets for strategy scholars. Responding to Bettis's (2012) call for better evidence, we provide specific estimates of the share of results that could be replicated; we evaluate evidence of "asterisk searching" and result nudging; and we calculate the extent to which published studies inflate true coefficients. In doing so, we complement Harrison's *et al*. (forthcoming) recent examination of meta-analyses that relate various factors to firm performance. Using a trio of methods to triangulate the aggregate publication bias, they find that reported effects are exaggerated between 0 and 30 percent, but they do not estimate the percentage of results that are likely to be replicated, nor do they study a broad sample of research in strategic management.

To conduct our analysis, we gathered data on estimates reported in 300 published articles in a random stratified sample from *Strategic Management Journal, Academy of Management Journal, Organization Science, Administrative Science Quarterly,* and articles in *Management Science* accepted by the strategy, entrepreneurship, or organization divisions. Our sample includes six articles from each journal for each year from 2003 to 2012 (a total of 60 per journal). To ensure that our results are not an artifact of the publication process, we gathered data from 60 additional proposals submitted

DOI: 10.1002/smj

to three prestigious strategy conferences. This conference series encourages proposals of early-stage, cutting-edge research; the review process does not prioritize significant findings (a majority of accepted papers do not include quantitative data); and about 60 percent of all proposals are accepted.

We estimate that between 24 and 40 percent of published findings based on "statistically significant" (i.e., $p < 0.05$) coefficients could not be distinguished from the Null if the tests were repeated once. Our best guess is that 29 percent of published results are non-replicable, of which about 70 percent should be interpreted to be 0. For the remaining 30 percent, the true B is not 0, but insufficient test power prevents an immediate replication of a significant finding. For these relationships, we estimate that the reported effects are inflated by 13 percent relative to the true effect.

We provide evidence that the region just past customary cut-off levels includes an abundance of false or inflated findings. By analyzing conference proposals, we demonstrate that a bias toward "significant" results is not simply a consequence of the review process: Even in early stage work, authors seem to filter results to find those that are statistically significant. In our conclusion, we suggest ways to reduce, if not prevent, scientific apophenia in strategic management.

EMPIRICAL METHOD

We use two approaches to analyze our sample of reported test statistics.

Method 1

Our first method builds on previous work that uses the distribution of test statistics to infer the number of false positives (true $B = 0$) and true results (c.f. DeLong and Lang, 1992; Simonsohn, Nelson, and Simmons, 2013). All of these studies are based on the recognition that distributions of test statistics carry clues about the true coefficients the empirical studies sought to test. When no true relationship exists (true $B = 0$), repeated testing creates a distribution of test statistics centered on 0. By definition, some estimates (5%) of this null relationship result in *t*-statistics greater than 1.96, but small *t*-statistics dominate the distribution. In contrast, repeated tests of a real relationship ($B \neq 0$) create a distribution of test statistics that is shifted away from 0. Each test statistic is a draw from this distribution. The percentage of these tests that exceeds 1.96 is defined as the "power" of the statistical test. More precisely, *statistical power* is the likelihood that a null hypothesis will be rejected when the relationship is in fact true. Statistical powers are further discussed and explained in Appendix S2, Part A.

Method 1 is designed to characterize our empirical sample of test statistics by comparing it to distributions simulated by combining tests of a false relationship ($B = 0$) and tests of a true relationship ($B \neq 0$) tested with a given statistical power. To avoid selection problems caused by "acceptance bias," we limit our main analysis to statistics with *p*-values below 0.05 (two-tailed). To avoid problems with comparing densities where few observations exist, we limit our analysis to statistics with *p*-values above 0.0001 (two-tailed). We use the Kolmogorov-Smirnov (KS) equality-of-distributions test to measure the degree to which simulated distributions match observed ones. This nonparametric test is based on the maximum distance between the lines of two cumulative distributions (Massey, 1951). Further details are provided in Appendix S2, Part B.

To find distributions consistent with the published test statistics in our sample, we conduct a grid search of mixtures of false positives and simulated statistics from tests with varying degrees of power. We first test 10 distributions created by combining 10 percent false positives mixed with distributions from tests with 10, 20, 30 … 90 percent power. Then, we test distributions with 20 percent false positives and 9 different levels of power. We continue until we have tested 90 combinations. The true distribution of published test statistics is likely to include tests with a continuum of statistical powers (Gelman and O'Rourke, 2014) and is fundamentally unknowable, but our method bounds this continuum. Our method cannot identify whether we have few false positives and many very weak results or many false positives and a few strong results, but it can estimate the feasible region for these mixtures. We thus provide insight on the likelihood that any result could be replicated.

To account for the dependence of results within any given study, we selected from each publication a single statistical model to evaluate. We gave priority to the model that tested all of the hypothesized relationships (generally referred to

DOI: 10.1002/smj

as the "full model"), or if this was lacking, we chose the most complete model. To account for the interdependence of estimates within these models, a bootstrapping method was used to randomly select a single test statistic for comparison. For each of the 90 simulated distributions we wished to analyze, we ran 500 KS tests using such randomly selected coefficients. Although computationally expensive, this method allowed us to use all of the relevant estimates while ensuring test statistics were independent within each KS test. Appendix S3 contains the STATA code used to perform the test.

Method 2

Our second method uses a posterior predictive check to evaluate the assumption that authors are reporting unbiased results (Gelman, Meng, and Stern, 1996; Rubin, 1984). Posterior predictive checks are usually grouped with Bayesian analysis, but are not fully Bayesian because they violate the likelihood principle. Gelman *et al*. (1996) and Rubin (1984) recommend posterior predictive checks be used, as we do, for testing the assumptions of models. As is usual in this method, we begin with a model of how the observed data were created that includes some assumptions about an unobserved parameter. In our case, this is the assumption that authors in our sample reported coefficients and standard errors that are unbiased by improper testing protocols such as fishing, motivated outlier removal, selective reporting, and so on. We then generate a predicted distribution using this model and compare it to our observed distribution. The greater the difference between the predicted and observed distributions, the more likely our initial assumption is false. We can also form, for any region of the *t*-statistic, an estimate of the number of results added or subtracted by deviating from proper testing protocols. Finally, we can estimate directly the probability that any finding will be significant in a single repeat test.

We assume that the coefficient values in future tests will be drawn randomly from $N(B_0,SE)$ and that standard errors will be drawn from a Chi Square distribution of the degrees of freedom reported in the research and scaled to reflect the reported standard error. Appendix S5 contains the STATA code used to perform the test. We generate a single random draw for each reported test statistic to generate a simulated sample, and repeat this process 1,000 times to generate an accurate 95 percent confidence interval for the expected density distribution of *t*-statistics from any single repetition of all of the studies in our sample.

One limitation with our second method is that we assume that our sample of research (and coefficients) is not biased by the submission or review process. If authors are less likely to submit papers with "nonsignificant" coefficients and/or journals less likely to publish them, then some of our results could be biased by this selection process. In one robustness test, we check this by using estimates from Method 1 to simulate an unbiased distribution.

Empirical sample

Table 1 reports descriptive statistics for the samples as described in the introduction. The data collection process is described in Appendix S2, Part C.

RESULTS

Method 1

Table 2 reports the mean *p*-value of the KS tests for each comparison of simulated and actual distributions. The confidence intervals on these estimates, reported in Appendix S4, are very small. Each column in Table 2 marks the percentage of the false results used in the simulated mixture and each row marks the power of the tests of true results that make up the balance of the mixture. Small values of *p* suggest that a particular test mixture (e.g., 40% false and the rest from a distribution generated by tests with 50% power) would be unlikely to produce a distribution like the one generated for reported test statistics. Darker shades associated with higher *p*-values suggest a better match between the simulated and observed distributions.

We find a range of simulated mixtures that could match our empirical distribution ($p < 0.05$). For example, we cannot rule out the possibility that the tests in the sample are formed from a 30 percent/70 percent mix of false positives and true results measured with 90 percent power. Such a mixture suggests that 37 percent of the coefficients would not be found to be significant in a single repeat test: 30 percent are actually false, and 10 percent of the 70 percent of true results would lose significance because the test has 90 percent

DOI: 10.1002/smj

Table 1. Descriptive statistics

	Published													
	All		AMJ		ASQ		MS		OS		SMJ		Conference proposals	
Article Statistics	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Number of hypotheses	4.87	2.72	5.45	3.01	5.40	2.84	3.90	2.39	5.12	3.00	4.48	2.00	3.63	1.57
Number of coefficients estimated	97.76	79.16	87.82	93.58	125.75	101.63	93.35	56.82	108.98	77.05	72.90	42.80	49.53	54.53
Number of hypothesized coefficients estimated	13.87	13.66	12.33	10.56	12.47	12.19	15.68	14.02	16.73	18.87	12.12	10.67	7.33	6.82
Tests per hypothesized coefficient	2.85		2.26		2.31		4.02		3.27		2.70		2.02	
Sample N	28,866	195,435	5,333	18,326	36,625	163,391	90,000	397,673	9,238	46,998	2,808	8,297	10,477	29,329
Sample N Median	840		1,020		1,146		855		1,029		483		462	
Significance cut off (two-tailed)	0.081	0.041	0.085	0.038	0.085	0.040	0.075	0.039	0.081	0.040	0.081	0.050	0.071	0.025

N = 300 manuscripts for published sample. 60 for each journal, and 60 for conference proposals sample.

power. In contrast, our results also suggest it is conceivable that the sample is made up of low powered tests (e.g., 60% power) and few false positives. In this case, 40 percent would not be replicated because the tests used in the sample only have 60 percent power. Considering the entire range of feasible mixtures, we estimate that between 36 and 52 percent of the coefficients in the range analyzed ($0.0001 < p < 0.05$) would not be judged "significant" in a single repeat test. For the conference proposals, 37–70 percent would not again have significant p values.

The above estimates are based only on analysis of the coefficients with p-values between 0.0001 and 0.05. If we wish to make estimates for all coefficients with $p < 0.05$, we must take into account the coefficient estimates where $p < 0.0001$ (34% of the total for published manuscripts, 32% for conference proposals). Estimates with such low p-values likely would be significant in a repeat test, and adding them to the analysis thus reduces the percentage of results we estimate would not be confirmed. After adding these additional tests into the denominator of our percentage calculation, we estimate that 24–34 percent of *all* published findings based on significant coefficients ($p < 0.05$) would not be corroborated by a single repeat test. We estimate that 25–52 percent of significant findings in our sample of conference proposals would not be significant in a repeat test. Our estimates represent a lower bound for the number of findings that would not be confirmed, we believe, because we only analyzed findings where $p < 0.05$, thereby excluding weaker findings reported as significant despite having $p > 0.05$.

Although we cannot rule out a number of possible mixtures, some match the empirical distribution more closely than others. For published results, the distribution created by mixing 30 percent false positives and the remainder from a true relationship tested with 80 percent power has the highest p-value ($p = 0.48$), suggesting this simulated distribution deviates least from the observed one. Using this to give a ball-park estimate, for all currently significant coefficients (i.e., all $p < 0.05$), 71 percent of coefficients would be confirmed as significant in a single repeat test of the studies (20% are lost because they are false; 9%, because of insufficient power). This "best" mixture also suggests the currently reported but low power results overestimate the magnitude of the true coefficients by 13 percent. This inflation factor is calculated as the ratio of the mean of

DOI: 10.1002/smj

Table 2. Results of KS comparisons of simulated and observed distributions

Power of tests of real results	Percent false positives										
	100	90	80	70	60	50	40	30	20	10	0
Analysis of published results in five strategy outlets											
10	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0.01
50	0	0	0	0	0	0	0	0	0.01	0.03	0.08
60	0	0	0	0	0	0	0	0.01	0.08	0.2	0.17
70	0	0	0	0	0	0	0.01	0.12	0.36	0.21	0.02
80	0	0	0	0	0	0	0.08	0.48	0.29	0.01	0
90	0	0	0	0	0	0.01	0.26	0.39	0.01	0	0
Analysis of results from conference proposals											
10	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0.01	0.02	0.03
30	0	0	0	0	0	0.01	0.01	0.03	0.06	0.12	0.21
40	0	0	0	0	0.01	0.02	0.05	0.12	0.25	0.41	0.53
50	0	0	0	0	0.01	0.05	0.15	0.34	0.56	0.56	0.31
60	0	0	0	0.01	0.03	0.13	0.34	0.61	0.57	0.23	0.04
70	0	0	0	0.01	0.07	0.25	0.59	0.62	0.22	0.03	0
80	0	0	0	0.02	0.13	0.42	0.68	0.29	0.04	0	0
90	0	0	0	0.03	0.16	0.48	0.47	0.07	0	0	0

Each cell reports average *p*-values of Koglomorov-Smirnov (KS) tests comparing 500 simulated distributions to the distribution of published tests statistics for hypothesized coefficients (with $0.0001 < p < 0.05$). Each column reports mixtures using a given share of false positives (true $B = 0$). Each row reports mixtures with simulations of tests of real results (true $B \neq 0$) with particular statistical power. For any shaded cells, we cannot reject the null (at $p < 0.05$) of a different distribution. Darker shaded cells reflect higher *p*-values—thus implying more similarity. The 95 percent confidence interval for the *p*-values is approximately 0.01.

the distribution of *t*-statistics with 80 percent power over the same distribution truncated at the critical value of 0.05. For the proposals submitted to the strategy conferences, the best result is created by mixing 40 percent false positives and the remainder from a true relationship tested with 80 percent power. For the entire region $p < 0.05$, this corresponds to 27 percent false positives, a confirmation rate of 65 percent, and an inflation of the true coefficients by 13 percent.

To confirm that our results are representative of strategy research and not biased by other types of publications in top strategy outlets, we reran the analysis on a subsample of the published research using only those papers members of the Strategy Research Initiative felt were "certainly" strategy. Results for the subsample closely match those for the full sample.

Method 2

In our second method, we assume that each individual report is correct and unbiased, and then evaluate whether, conditional on this assumption, the distribution of observed test statistics falls within the 95 percent confidence interval for distributions of simulated repeat studies. Figure 1(a) shows the resulting simulations for published manuscripts. The graph reveals a region ($2 < t < 3$) where the observed distribution passes outside of the 95 percent confidence interval. The area measuring these unexpected estimates represents about 21 percent of all hypothesized coefficients where $2 < t < 3$. The graph also exhibits a region of low test statistics ($t < 0.75$) where observed results are below the 95 percent confidence interval. This is consistent with the premise that some papers with low test statistics are selected out during the review process or never submitted by authors for review. Another region with higher *t*-statistics ($t > 3.75$) also falls near or below the bottom of the 95 percent confidence interval. This is because the high density of test statistics near 2.0 implies the existence of more powerful tests than the observed sample actually includes.

In conducting the above analysis, we assume that the review process does not bias our sample of manuscripts. Were significant coefficients (i.e.,

DOI: 10.1002/smj

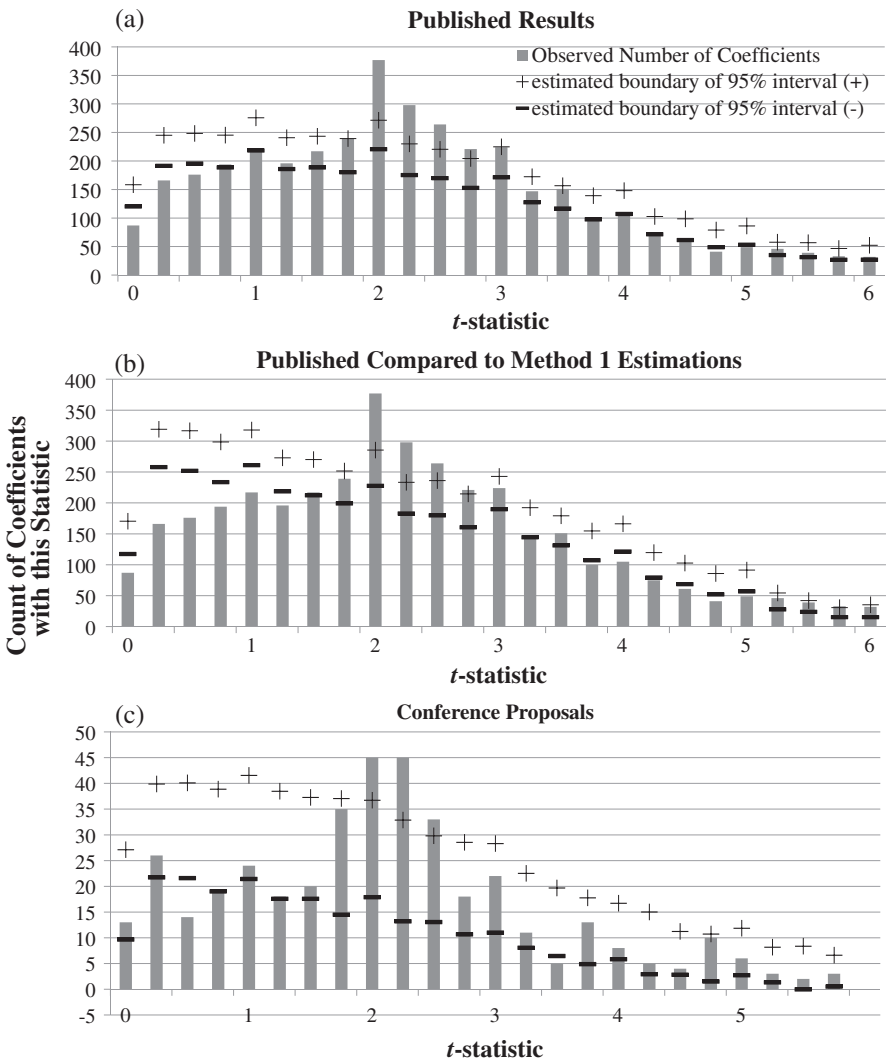

Figure 1. Method 2—Distributions of actual vs. expected *t*-stats for hypothesized coefficients. The three histograms report the distribution of test statistics in published papers in top outlets for strategic management (a, b), and those from conference proposals (c). Also displayed is the 95 percent confidence interval for the histogram were all studies to be run again on a separate but equivalent sample drawn from the same population. The 95 percent confidence interval is calculated using either observed coefficients and standard errors (a, c) or coefficients and standard errors created by the best matching distribution found using Method 1 (see Table 2)

$p > 0.05$) to be differentially rejected, our results would be distorted by a failure to observe estimates that were removed in the review process. As a check on the effect of such sample-selection, we used the estimates from Method 1 to simulate possible coefficients censored by the review process. Figure 1(b) shows the observed results compared to a simulation based on the best mixtures found in Method 1. Two regions are evident. First, the simulation suggests the density of small test statistics ($0 < t < 1.5$) is reduced by at least 30 percent. In other words, in this region selection by authors or reviewers reduces the number of "nonsignificant" estimates. This region represents the effect of journals selecting for significant results or the result of authors leaving nonsignificant findings in file drawers. In contrast, the region just past usual significance cut-offs ($2 < t < 3$) has at least an 18 percent unexpected abundance in test statistics. This profusion of results near significance cut-offs is likely a product of fishing or *p*-hacking.

DOI: 10.1002/smj

So far, our analysis has not distinguished between the effect of the review process and author submission decisions. Analysis of our data on proposals provides some evidence in this regard. If authors are producing manuscripts with many significant results, but a review process emphasizing significant results is selecting for those that meet critical value thresholds, then the test-statistics for the conference proposals should include more nonsignificant results and fewer coefficients just past significance cut-offs. A visual comparison of the distributions shown in Figure 1(a, c) suggest that this is not obviously the case. Quantitative analyses of these distributions conducted in Method 1 suggested they are likely to be comprised of a similar mix of false and true results. Furthermore, Method 2 analyses of published papers and conference proposals suggest that a similar number of coefficients have been fished or hacked into place.

Finally, we use Method 2 to check Method 1's estimates of the potential for a single future study to confirm published results. We can compute this directly by simply counting the percentage of the currently "significant" coefficients that land in the $p < 0.05$ range in each repeat simulation. Such an approach makes no assumption about selection by reviewers or journals, and it makes the generous assumption that all existing coefficients and standard errors are unbiased. Using this method, we estimate that between 38 and 40 percent of the currently "significant" results would not be confirmed by a single repeat test. This is a slightly more pessimistic result than the estimate generated with Method 1 (25–34% failed confirmations). We suspect that the difference is caused by Method 1's inability to create a simulated distribution that matches the extraordinarily high mass of results with *t*-statistics close to 2.0. These barely significant results have only a 50 percent chance of being confirmed.

Future research

In future work, we plan to analyze differences among the journals. We separately analyzed each journal and found evidence of apophenia in each. We do not report results for our separate analyses here because proper interpretation will require a lengthy and distinct study. A finding that one journal has stronger results than another may be misleading if journals also differ with respect to norms about what constitutes a hypothesis. Some journals include more hypotheses per paper, and some of these serve as a starting point by confirming well-established relationships. Other journals limit the hypothesized tests to those that are most speculative.

We also plan to examine if scientific apophenia is more common when researchers analyze certain types of data. Correcting known biases in data, for example, may give authors an opportunity to experiment with alternative econometric specifications and lead to post-hoc rationalizations of the "best" results. To the degree that authors can choose which method to report, there is a greater opportunity for scientific apophenia. We also plan to investigate how mistaken inference from poor identification is associated with misinterpretation of test statistics.

DISCUSSION AND REMEDIES

In this article, we answer Bettis's (2012) call for a baseline measure of the scale of the problem in the literature on strategic management. Some of the results are encouraging. A majority of published findings, we find, probably document a real effect (the true $B \neq 0$), and because of the veracity and strength of tested relationships, more than 60 percent are likely to be confirmed by a single repeat test. Yet, there is also reason for concern. We estimate that between 24 and 40 percent of published findings based on "statistically significant" (i.e., $p < 0.05$) coefficients could not be distinguished from the Null if the tests were repeated once. Our best guess is that for about 30 percent of these nonconfirmed results, the coefficient should be interpreted to be 0. Furthermore, published estimates of real results probably inflate the magnitude of the true effects by 13 percent. No silver bullet exists to fix these problems, but partial solutions are readily available.

Education could make scholars aware of simple practices that can reduce the chance of finding a false positive. Individual researchers can apply a textbook strategy from the field of data mining by randomly splitting their data in two before beginning statistical analysis (Bettis, 2012; Han, Kamber, and Pei, 2006). If an effect is random, then it will be unlikely to appear in the second, reserved test sample. This method requires samples of sufficient size to allow sufficient power in each half sample to conduct a meaningful statistical analysis.

DOI: 10.1002/smj

We do not think the value of split samples should bar authors from using small samples to analyze interesting questions, but we do think that the inability to conduct such a simple test should encourage caution in making inferences. Discussions and presentations of all empirical studies should reflect the limits of the studied sample.

Education could also help clarify the meaning of test statistics. Many scholars seem to believe that $p < 0.05$ is both necessary and sufficient to make a finding "meaningful." For example, during a presentation showing the prevalence of false positives from repeat searches of random data, one top strategy scholar exclaimed: "but if [$p < 0.05$], those are real findings" – ignoring the fact that five percent of a distribution of *p*-values of a random relationship will, *by definition,* fall below the significance cut-off. Recognizing that the dichotomous treatment of test statistics distorts the meaning of *p*-values could also encourage authors to be more thoughtful about their interpretation of test statistics. There is little inferential difference between a coefficient with a *p*-value of 0.04 and one of 0.06, yet scholars often infer the former as providing "support" and the latter as providing "no evidence."

An awareness of the history of *p*-values might help deflate their swollen stature and encourage more judicious use. We were surprised to learn, in the course of writing this article, that the $p < 0.05$ cutoff was established as a competitive response to a disagreement over book royalties between two foundational statisticians. In the early 1920s, Kendall Pearson, whose income depended on the sale of extensive statistical tables, was unwilling to allow Ronald A. Fisher to use them in his new book. To work around this barrier, Fisher created a method of inference based on only two values: *p*-values of 0.05 and 0.01 (Hurlbert and Lombardi, 2009). Fisher himself later admitted that Pearson's more continuous method of inference was better than his binary approach: "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects [null] hypotheses; he rather gives his mind to each particular case in the light of his evidence and ideas" (Hurlbert and Lombardi, 2009: 316). A fair interpretation of this history is that we use *p*-values at least in part because a statistician from the 1920s was afraid that sharing his work would undermine his income (Hurlbert and Lombardi, 2009). Following Fischer, we recommend that authors report *p*-values and refrain from emphasizing thresholds. This will allow us to more easily interpret evidence on a continuum and in the context of previous findings.

We believe that increased reporting of research methods data and models could help reduce the number of false positives. Simmons *et al.* (2011) suggest that authors report data collection plans, experimental conditions tested, variables measured, observations eliminated, and specifications used. Bettis (2012: 112) also suggests "reporting the actual number of tests conducted." Such reporting could make scholars more mindful of their own practices, and allow reviewers and readers to better assess the veracity and power of reported results. Disclosure of data will only reduce apophenia if scholars are properly incentivized to reanalyze published studies (Simmons *et al*., 2011). The upcoming replication special issue in this journal is a step in this direction.

Obviously, reporting and disclosure requirements will need to be advocated and enforced by academic journals. Radically new procedures for reviewing could also be considered. Economist Robin Hansen (2014), for example, has proposed, "results blind reviewing" in which results are withheld until after reviewers have accepted or rejected the paper. Alternatively, authors could be required to register their hypotheses before initiating their research. In 2013, scholars from numerous disciplines began a campaign to encourage just that (Chambers and Munafo, 2013). We would do well to learn from our peers in neighboring disciplines.

Changes in reviewing and publication procedures will have limited effect if perceptions of what qualifies as "good" research go unchallenged. Many educational institutions reward scholars based on the number, rather than the rigor, of their publications. Theory and "thought leadership" is often valued more than empirical analysis, while replication studies are viewed as marginal contributions. Yet, without replication, we cannot hope to understand the power and usefulness of the very theories we claim to value. Currently, the impact of our ideas is not limited by their lack of novelty, but by their lack of certainty. Scholars and practicing managers will make better use of research that is precisely defined and definitively demonstrated.

As expressed by Nobel Laureate Richard Feynman: "The first principle [of good science] is that you must not fool yourself—and you are the easiest person to fool." In this spirit, we call on strategy scholars, reviewers, and journal editors

DOI: 10.1002/smj

to reconsider what it means for something to be known. In our experience, a single positive finding is often interpreted as proof of a theory, when, in fact, it usually should not be considered as anything more than intriguing. An estimated coefficient is a single statistic from a distribution. It takes many estimates to understand the true nature of that distribution. It is for this reason that separating false and true results generally requires thoughtful analysis and numerous replications. It also requires scholars, journals, and institutions to value their ideas enough to be willing to see them rigorously tested.

ACKNOWLEDGEMENTS

We thank Rajshree Agarwal, Richard Bettis, Seth Carnahan, Cristian Dezso, Daniel Feiler, David Kirsch, Daniel Malter, Timothy Simcoe, Bill Simpson, Paul Wolfson, Bennet Zelner, and seminar participants at the University of Maryland Strategy Brownbag for helpful comments. Chenqi (Bill) Huang, Ying Geng, Siddhartha Sharma, Baljir Baatartogtokh, and Justin Frake provided excellent research assistance. All errors are our own.

REFERENCES

Bacon F. 1620. Novum organum, aphorism 45.

Bettis RA. 2012. The search for asterisks: compromised statistical tests and flawed theories. *Strategic Management Journal* **33**(1): 108–113.

Chambers C, Munafo M. 2013. Trust in science would be improved by study pre-registration, *theguardian.com*, Wednesday 5 June 2013.

DeLong JB, Lang K. 1992. Are all economic hypotheses false? *Journal of Political Economy* **100**: 1257–1272.

Denton FT. 1985. Data mining as an industry. *Review of Economics and Statistics* **67**: 124–127.

Feynman RP. 1985. *Surely You're Joking, Mr. Feynman!: Adventures of a Curious Character*. WW Norton & Company: New York.

Gelman A, O'Rourke K. 2014. Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics* **15**(1): 18–23.

Gelman A, Meng X-L, Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica* **6**(4): 733–760.

Han J, Kamber M, Pei J. 2006. *Data Mining: Concepts and Techniques*. Morgan Kaufmann: Waltham, Massachussets.

Hanson R. 2014. Overcoming bias. Available at: http://www.overcomingbias.com/2010/11/results-blind-peer-review.html. (accessed 29 June 2014).

Harrison JS, Banks GC, Pollack JM, O'Boyle EH, Short J. (forthcoming). Publication bias in strategic management research. *Journal of Management* .

Hurlbert SH, Lombardi CM. 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici* **46**(5): 311–349.

Massey FJ Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association* **46**(253): 68–78.

Merriam-Webster. 2014. Available at: Merriam-Webster.com. (accessed 29 June 2014).

Rubin DB. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**(4): 1151–1172.

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**(11): 1359–1366.

Simonsohn U, Nelson L, Simmons J. 2013. P-curve: a key to the file drawer. *Journal of Experimental Psychology* **143**(2): 534–547.

Stanley TD. 2005. Beyond publication bias. *Journal of Economic Surveys* **19**(3): 309–345.

Tversky A, Kahneman D. 1973. Availability: a heuristic for judging frequency and probability. *Cognitive Psychology* **5**(2): 207–232.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix S1. STATA code for simulation 1
Appendix S2. Statistical Power, Distinguishing Distributions & Data Collection Procedures.
Appendix S3. Method 1 STATA code
Appendix S4. Method 1 confidence interval calculations
Appendix S5. Method 2 STATA code