TESTS FOR IDENTIFYING "RED FLAGS" IN EMPIRICAL FINDINGS: DEMONSTRATION AND RECOMMENDATIONS FOR AUTHORS, REVIEWERS AND EDITORS

DONALD D. BERGH
Daniels College of Business
The University of Denver
2101 S. University Blvd.
Denver, CO 80208
dbergh@du.edu

BARTON M. SHARP
Department of Management
Northern Illinois University
Barsema Hall
DeKalb, IL 60115
bsharp1@niu.edu

MING LI
Hull University Business School
Cottingham Road
Hull, Yorkshire
HU6 7RX
United Kingdom
Lilyliming@hotmail.com

The authors gratefully acknowledge the contributions of Herman Aguinis in contributing toward the development of this paper.

TESTS FOR IDENTIFYING "RED FLAGS" IN EMPIRICAL FINDINGS: DEMONSTRATION AND RECOMMENDATIONS FOR AUTHORS, REVIEWERS AND EDITORS

Abstract

High profile article retractions, survey results indicating falsification of data, and evidence of mistaken findings raise concerns that problematic empirical research has found its way into the management field's literatures. To help safeguard the field against such vagaries, the authors describe three tests that can be applied to most empirical articles to assess the accuracy of the reported findings. They also demonstrate how the tests uncover reporting anomalies using a retracted article as an example. The results identify numerous irregularities which would have raised "red flags" had the tests been applied to the article while it was under review. The authors offer several recommendations to help protect the trustworthiness of management research.

Keywords: Verification, replication, reproducibility, academic misconduct

> *Prof. Dr. Lichtenthaler informed the Rector of the University of Mannheim that he wants to leave the University of Mannheim on March 31, 2015. The state of Baden – Württemberg has agreed with his wishes.* Press Release, October 2014, Universitaet Mannheim (http://www.uni-mannheim.de/1/presse_uni_medien/pressemitteilungen)

As of April 2016, 16 articles authored by Ulrich Lichtenthaler have been retracted from the management field's top academic journals including the *Strategic Management Journal*, *Academy of Management Journal*, *Organization Science*, *Research Policy*, the *Journal of Management Studies*, and others (http://retractionwatch.com/the-retraction-watch-leaderboard/). Such retractions may not be surprising. A survey of management faculty at research-intensive institutions reports evidence of data fabrication, finding falsification, and plagiarism (Bedeian, Taylor & Miller, 2010) while other studies document that more than 20 percent of reported significant statistical findings may be inaccurate (Bakker & Wicherts, 2011; Goldfarb & King, 2016; Nuijten, Hartgerink, Assen, Epskamp & Wicherts, 2015). Overall, instances of retractions, possible scientific misconduct, and honest mistakes pose a worrisome threat to the trustworthiness of accumulative knowledge – the cornerstone of effective evidence-based management (Kepes, Bennett, & McDaniel, 2014) - and raise concerns about the validity of the field's theory development and recommendations for practice.[1]

Unfortunately, few barriers are in place to keep problematic studies from slipping into the field's knowledge base. Schminke, for example, noted that "…we have no formal, mandatory audit process…I have never once been asked…to show my data, much less the records involved in collecting and assembling those data. In my tenure as associate editor of the *AMJ*, and more

[1] Such practices may also have implications for personal relationships. Holger Ernst, Ulrich Lichtenthaler's coauthor, was formally reprimanded for "having not sufficiently reviewed his work for mistakes, and the commission judges this behavior as severe scientific misconduct….by neglecting this duty Professor Ernst bears shared responsibility for the errors occurring in the joint publication" (http://retractionwatch.com/2015/09/29).

recently at *Business Ethics Quarterly*…I never had even a single reviewer request access to data" (2009: 590). More generally, the management field lacks a mechanism for routinely assessing the trustworthiness of the scientific knowledge it produces (Kepes *et al*., 2014: 448), reviewers and editors often miss even the most egregious of methodological flaws (e.g., Bohannon, 2013; Godlee, Gale & Martyn, 1998; Schroter *et al*., 2008), and replication studies tend to focus less on discrepant findings and more on differences in study features (Hubbard, Vetter & Little, 1998). Consequently, at present, the field's empirical foundation and its recommendations heavily depend on author integrity and complete accuracy in all data reporting and interpretation.

These research norms raise questions: How many errant or fraudulent conclusions are we willing to tolerate in our literature? What are we willing to do to screen them out? We submit that a reformulation of disclosure and publication requirements is needed to safeguard the trustworthiness of reported empirical findings in management research. Such revisions should include objective and independent tests to confirm the accuracy of reported results. In this article, we describe three tests that can be applied to verifying the findings of most empirical studies in management research. We then use a retracted article to demonstrate how the tests uncover reporting irregularities. We close with recommendations for how authors, reviewers and editors can work together to protect the body of empirical work in management.

Overall, the tests described in this article represent one step toward proactively safeguarding the trustworthiness of knowledge rather than leaving the field's empirical base vulnerable to exploitation and error. We recognize that the tests do not apply to all articles, and have limitations themselves, but nonetheless installing a mechanism for assessing the accuracy of reported findings seems a necessary stage in the review process to help ensure the credibility of the field's body of empirical findings.

THREE TESTS

We searched the management literature, as well as psychology, economics, and sociology, to identify objective tests that can be used by an independent party to assess the accuracy and validity of reported empirical findings. We used two screens to identify all possible tests including: (1) those that do not require access to the authors' original data but can use information reported in a manuscript as input instead and (2) appear in peer-reviewed journal articles. Tests passing these screens could be applied to the largest possible scope of studies, would be accessible to the highest number of possible testers, and had met the standards of peer review.[2]

Three tests were identified. One examines the congruence of reported and reproduced test statistics (*t, f, z*), degrees of freedom and *p* significance levels; another draws upon a simulation-based verification methodology to compare reported and expected significance levels; and a third uses matrices of reported descriptive statistics of a study's data to retest the study's reported models. Table 1 presents each test and its respective advantages and disadvantages.

---Insert Table 1 about here---

Test One: Congruence of reported test statistics

This first test has recently been applied in psychology journals (e.g., Bakker & Wicherts, 2011; Nuijten *et al*. 2015) to identify cases where published findings may contain errors in the reporting of statistical results. In general, the test evaluates the level of consistency of statistical results associated with null hypothesis significance testing (NHST), whereby reported *p*-values

[2] We thank an anonymous reviewer for identifying additional tests that can be applied to assessing a study's findings. These assessments tend to require complete data sets for the analyses. We describe some below in the Discussion section.

are considered relative to their accompanying test statistics and degrees of freedom (*df*). More specifically, Bakker and Wicherts (2011: 668) describe the tests as follows: "We gleaned from each article the test statistics, *df*, and *p* value…we recalculated the *p* value on the basis of the reported test statistic and *df* and compared these values with the reported *p* values. We considered a reported *p* value to be incorrect if it differed from our recalculated *p* value." Given that the perceived support, or lack thereof, for a theoretical hypothesis is generally based on the reported *p* value, a difference between what was reported and what the *p* value should have been based on the underlying statistics could affect the substantive conclusions and contributions of the focal study.

This approach is direct, straightforward, and allows "apples to apples" comparisons of reported statistical significance *p*-values for control, independent, moderating, and mediating relationships. Further, the tests can be applied to large samples through using software packages that read entire articles as input; for example, the recently developed procedure *statcheck* within the R package (version 1.0.1; Epskamp & Nuijten, 2015) can extract statistical results from PDF or HTML files and recalculate *p*-values based on the reported statistical results and their degrees of freedom. The test does suffer from some drawbacks, namely that it requires a complete disclosure of essential statistics. For instance, reporting only coefficients and *p*-values is insufficient to permit the evaluation, as the tests also need either standard errors and parameter statistics (*t, f, z*) or the degrees of freedom. Further, the test identifies only the congruence of the reported significance levels and cannot ascertain whether authors misreported or distorted their statistics in other ways beyond simply misstating the statistical significance of particular coefficients. The test is also vulnerable to the clarity of author reporting. Decisions such as using one dataset for one table and another dataset for others, copy and paste errors, and the use of one-

tail or two-tails test could not be detected unless disclosed (Bakker & Wicherts, 2011). Finally, the test cannot provide insights into the size and direction of coefficients as it focuses on significance levels instead.

Test Two: Simulation-based verification

A recent *Strategic Management Journal* article by Goldfarb and King (2016) applies a simulation-based test to estimate how many coefficients may be over- or under-stated relative to an expected "true" effect size. This approach involves several steps: (1) developing a model of observed data and an assumption about an unobserved parameter, where authors may have reported coefficients and standard errors that are potentially biased due to data manipulation, selective reporting, data snooping and others (see Bettis, 2012); (2) creating a predictive distribution for comparisons with the observed distribution; (3) using coefficient ranges to estimate the number of results relative to an expected level and (4) estimating the probability that any finding will be significant in a single repeat test. The underlying assumption is that "coefficient values will be drawn randomly from N(B0,SE) and that standard errors will be drawn from a chi square distribution of the degrees of freedom reported in the article and scaled to reflect the reported standard error…[where they] generate a single random draw for each reported test statistic to generate a simulated sample, and repeat this process 1000 times to generate an accurate 95 percent confidence interval for the *t*-statistics from any single repetition of all of the studies in [a] sample" (Goldfarb & King, 2016: 170).

More simply put, this procedure simulates what would happen if the published research were to be repeated numerous times with each repetition being done with a new random draw of observations from the same underlying population. The test results allow researchers to

characterize the stability or generalizability of published findings by answering the question: How likely is it that we would get the same results on a different sample from the same population? This method allows us to detect cherry-picking of samples or models even when the published descriptions of the data and results are perfectly accurate.

There are characteristics of the simulation method which limit its usefulness in detecting errors or malfeasance in published research. First, since the procedure is predicated on comparing the count of coefficients which fall into a given range of *t*-statistics relative to how many would be expected to if the regressions were repeated multiple times, a large number of coefficients are required to get meaningful results. Even in fairly expansive articles, including the replication examined below, the total number of coefficients would likely be too small to provide meaningful count data. That problem is compounded by the likely nature of the errors or malfeasance. If authors were to cherry-pick data to fit their theoretical agenda, they would primarily be interested in selecting data and models which gave them the desired results on hypothesized coefficients. They would have no incentive to bias the results on control items. Such a practice would exacerbate the small-numbers problem when applying the technique to one or even a small set of articles. For example, Goldfarb and King used their procedure to characterize the findings on approximately 4161 hypothesized coefficients across 300 published works. With an average of fewer than 14 hypothesized coefficients per article in their sample, there simply are not enough coefficients to calculate meaningful count data based only on hypothesized relationships from a single article. An alternative would be to include all coefficients, hypothesized or not, from the focal article. The problem, however, is that the control coefficients, which are more likely to fall into the "correct" range of *t*-statistics (since

there is no incentive for them to be biased), could mask errors or bias in the hypothesized coefficients.

A second shortcoming of the Goldfarb and King method is that it does not provide specific insight into which particular coefficients may have been misstated or inflated. While it can characterize the amount of potential malfeasance in a population of published research, it cannot pinpoint whether the statistical evidence regarding any particular theoretical hypothesis should be called into question. It is unable to isolate the precise coefficients which may be under- or over-reported within a population of studies.

Despite these limitations, it is valuable to include the Goldfarb and King technique in our catalog of tools for detecting errors or malfeasance as it could be extremely useful in detecting problematic patterns within particular bodies of research. For example, if questions were to arise about a given author's work, the Goldfarb and King method could be applied across their body of published articles to test for any systemic problems. Similarly, it could be applied across the body of research in a given theoretical area to possibly help explain inconsistent results (due, perhaps, to some authors cherry-picking results where others do not).

Test Three: Verification based on matrices of descriptive statistics

A final test for verifying study findings is to re-run a study's reported regressions using data derived from the published descriptive and correlational statistics. These recreated regressions can then be compared to the reported findings. Since the early 1980s, statistical packages such as SPSS have allowed researchers to create matrices of a study's variable means, standard deviations, correlations, and sample sizes which could then be analyzed as substitutes for the original raw data. Assuming all descriptive statistics are reported fully and accurately, these

analyses produce the exact same findings as regressions run on the original data (see Shaver, 2005; Boyd, Bergh and Ketchen, 2010, for illustrations within the management literature). To date, many other statistical packages including Stata and SAS also offer such a function.

This method of verifying published results by using the reported descriptive statistics and correlations to re-create a statistically equivalent dataset has a number of advantages. First, it is relatively easy, straightforward, and accessible to anyone with most major software packages that have built-in functions that take matrices of descriptive statistics as inputs to recreate the data. From that point on, the regressions can be run just as if the researcher had the original dataset. Second, this approach can effectively detect a number of different errors or misstatements. A mismatch between the coefficient sign and significance reported by an author and those obtained by running regressions on the recreated dataset would indicate either that (1) there was an error or typographical mistake in the published tables of descriptive statistics and correlations; (2) there was an error or typographical mistake in the published regression results; (3) authors chose to falsify results by reporting a coefficient sign or significance different than that which resulted from their regressions or (4) the regressions were run on a dataset that differed in some way from that described in the tables of means, standard deviations, and correlations, such as when an author might run regressions on a cherry-picked subsample of the original data in order to snoop for significant findings (e.g. Bettis, 2012).

This method, however, is not without limitations. For one, it offers no insight into which of the aforementioned problems might be in effect. The results can suggest reason for skepticism, but offer no specificity as to why. For another, this method would not detect a situation in which an author carefully selected observations that would lead to their desired empirical results and then reported both the descriptive statistics and regression results based on

that selected sample. This method is also limited to verifying models for which all predictor variables are explicitly included in the tables of descriptive statistics. The dataset recreated by the procedure is statistically equivalent to the data described by the means, standard deviations, and correlations, but the individual variable values in a given observation are meaningless. As a result, we cannot use those values as the basis for calculated variables such as multiplicative interaction terms. Therefore this method cannot be used to verify models with interaction terms, transformed variables, or squared terms unless those calculated variables are included explicitly in the descriptive statistics. We cannot tell based on this matrix-based verification whether the published results truly reflect a phenomenon in the underlying population, or if the results are an artifact of the particular sample drawn (even if the sampling was done honestly). This particular shortcoming of the matrix-based verification procedure is the biggest strength of the simulation method applied by Goldfarb and King (2016). Their procedure simulates what would happen if the published research were to be repeated numerous times, with each repetition being done with a new random draw of observations from the underlying population.

DEMONSTRATION

To illustrate how the tests work and the findings that they produce, we applied each to an article authored by Lichtenthaler and Ernst (2012, hereafter referred to as L&E) which was originally published in the *Strategic Management Journal* but subsequently retracted "at the authors' request due to material technical errors in the article…which have rendered many of the article's conclusions incorrect" (*Strategic Management Journal*, 2012: 1341). We selected this article to demonstrate how the three tests would have detected these "material technical errors." Our purpose is not to highlight the article, or to offer any generalizations about the body of empirical

findings in management research, but instead to show how the tests detect irregularities in findings and show researchers what to look for when conducting them.[3]

Overview of L&E (2012)

L&E (2012) examine whether "a firm's product development processes and technology licensing processes complements rather than substitutes in knowledge exploitation" (page 514). They offer three hypotheses that relate interactions of product development and technology licensing processes to firm revenues, licensing performance, and a firm's overall performance. Their study's data include semi-structured interviews with "45 R&D, innovation, marketing and business development experts in 30 firms from the automotive/machinery, chemical/pharmaceutical, and semiconductors/electronics industries [and]… a survey of the 300 largest firms" in those industries (2012: 520). They acknowledge that their data were also included in an earlier study, though the present study examined different variables. Their reported coefficients from reliability and validity tests meet conventional standards.

L&E (2012) report a correlation matrix (without the interaction terms) and unstandardized regression coefficients with standard errors. The findings from regression analyses are used to suggest partial statistical support for the first hypothesis and complete support for the second and third. These findings are augmented with supplemental slope analyses, additional exploratory regression analyses and split sample re-tests. Perhaps in an additional effort to garner credibility, the reference section includes four previous articles by Lichtenthaler, one by Ernst, and three by the respective editor. Overall, the authors conclude that "the data have emphasized that the identification of licensing opportunities strengthens the

[3] The full syntax of all tests conducted for this article are available upon request

positive effects of product development, whereas the commercialization stage does not significantly interact with product development…[and] has deepened our understanding of the intellectual property route to technology leveraging by means of licensing…has important managerial implications…[such as] most firms' traditional focus on product development may be insufficient" (page 530).

Test One: Findings from the Statistical Congruence Tests

Two coders independently collected the reported coefficients (*b*), standard errors (*se*), observations (*n*), number of variables (*k*), and degrees of freedom (*dfs)* for the variables in 29 analytical models reported in L&E.[4] Using Excel software, they each recalculated the statistical significance levels (*p*-values) for the *t*-values (=*b*/*se*) at their calculated *df* values and compared the 373 recalculated *p*-values in all 29 models to the reported *p*-values. The coders' initial findings agreed in 98 percent of the cases (365 of 373 *p*-values). The differences were due to entry errors which were subsequently resolved and 100 percent agreement in the findings was reached.

The re-test results for all coefficients in 29 models reported in L&E's study are presented in Table 2. First, all recalculated *p*-values were larger (less significant) than the originally reported *p*-values. Second, 28 of the 29 analytical models contained at least one non-verifiable result, and up to 40 percent of the variables in a given model had reported significance levels which were different from those we recalculated from the reported test statistics. In total, 77 *p*-values (21 percent of total 373 reported *p*-values) were discrepant between recalculated and reported *p*-values.

[4] One of the two coders was not an author. This coder was presented with the L&E article and asked to conduct the analysis independently.

---Insert Tables 2 and 3 about here---

Table 3 reports the results of re-testing the hypothesis coefficients. Fifteen recalculated *p*-values were different from reported *p*-values (as highlighted in bold font in Table 3). None appear to be due to rounding errors, all initial results in favor of the authors' hypotheses were reversed, and supported hypotheses lost empirical support in the recalculation. Overall, 65 percent (15 of 23) of the models which tested hypotheses report statistically significant *p*-values that could not be reproduced, and their supported hypotheses and conclusions from additional exploratory regression analyses and split sample re-tests lost empirical support. This relatively simple test indicated multiple "red flags" in the L&E article.

Test Two: Findings from the Simulation-based Approach

Two coders independently constructed and compared a data matrix that was to be used as input into the analytical procedures reported in Goldfarb and King (2016). The coders' findings were identical: The data values in the input matrix were exactly the same with one another as well as the data values reported in the L&E article. The analytical procedure used was double-checked to ensure that it was identical to the syntax published in an online supplement to the Goldfarb and King (2016) article.

The simulation technique uses characteristics of the *t*-statistic distribution to estimate the extent to which published regressions represent results that would be obtained by repeated study of the underlying population. Although this method is generally more suitable for testing multiple studies with large numbers of regression coefficients, it can be also applied to examine evidence of one article in a more limited fashion. Goldfarb and King (2016) report the *t*-statistic distribution for only those coefficients involved in hypothesis testing, since those are the

coefficients most likely to be biased or cherry picked by authors. Because they were using a large sample of articles (n = 300), they had enough such coefficients to make it statistically meaningful. Since our study endeavors to simply demonstrate the techniques on only one article, there are relatively few hypothesized coefficients to use as inputs into the Goldfarb and King algorithm. In an attempt to have a large enough number of coefficients to make their count-based analysis meaningful we included all 373 coefficients from the L&E article – spanning controls, independent and moderating variables, with no specification made for hypotheses.

---Insert Figure 1 about here---

The chart in Figure 1 shows how many coefficients from the L&E article were reported to be within a given range of t-statistic, compared to how many would be expected to fall within each range if the regressions were repeatedly re-run on new samples drawn from the same underlying population. The vertical dashed line denotes roughly the t=1.96 level, or the breakpoint between p<0.05 and p>0.05. To point out one example, the Figure indicates that there were ten coefficients from the results published by L&E which had a reported t-statistic of 1.9. The upper and lower confidence intervals are based on the results that would be expected if the same regressions were conducted 1000 times with each iteration using a new draw from the underlying population described by the reported statistical results. In this case the interval indicates that there is a 95% chance that the number of coefficients with a t-statistic of 1.9 should fall between four and 15. The fact that the actual number of reported coefficients with that t-statistic is within the bounds of the confidence interval suggests that those particular results are repeatable and generalizable to the population rather than being artifacts of decisions made by the authors.

Any interpretation of the results of a Goldfarb & King (2016) analysis applied to a single article must be considered carefully, as the relatively small number of coefficients leads to a lack of statistical power in the simulation. However in our Figure 1, which shows the results of applying the Goldfarb & King (2016) method to the L&E article, we can still see an example of the kind of result that would raise concerns in a more robust setting. Based on the simulation of re-running the regressions with 1000 unique draws of observations from the underlying population, there is a 95% chance that the number of coefficients with a *t*-statistic of 3.7 (corresponding to a significance of $p<0.001$) would be between zero and four. In the results reported by L&E there were actually five coefficients with that particular *t*-statistic. If such a result were found across multiple articles with a larger total number of coefficients and thus more power, it might suggest that the authors had cherry-picked models, samples, or results such that the reported results indicate more highly significant coefficients than what would be expected if the study were repeated with a new sample from the same population.

It would be difficult to draw any such conclusion from this one demonstration, both because of the lack of statistical power as well as the fact that there also appears to be an over-reporting of coefficients with t-statistics of 0.3 and 1.0 (both of which correspond to insignificant *p*-values). A more striking example of what a researcher should watch for when applying this method is available in Figure 1, Chart A of Goldfarb and King (2016: 173). Based on the 300 articles in their sample, there seems to be a significantly higher number of reported coefficients in the *t*-statistic range from 2-3 than we would expect to see if those models were re-run with new samples drawn from the same distribution, along with a correspondingly lower number of reported coefficients in the t-statistic range from zero to one.

Test Three: Findings from the Verification Based on Matrices of Descriptive Statistics

As with the first test above, two coders again independently conducted the analysis. Each also used a different statistical software package (Stata and SPSS). In both cases, the correlation matrix, means, standard deviations and sample sizes were used to create data matrices which were subsequently used to retest the base regression models reported in L&E. The regression analyses conducted by the two coders produced identical results.

---Insert Table 4 about here---

Table 4 presents the findings. Unfortunately, L&E did not disclose the interaction terms within their correlation matrix, so we were only able to test the base models and not those containing the product terms. Even so, our findings reveal numerous discrepancies between the reported and reproduced values (again highlighted in bold font) that raise questions about the accuracy and validity of the models in general. Indeed, none of the six base models could be reproduced in its entirety; in most cases, coefficients reported as significant were not confirmed in our tests. Although these re-tests cannot be applied to the product-terms, the consistent non-duplication of findings is compelling evidence of "red flags" consistent with the author's acknowledgment of "material technical errors."

DISCUSSION

Recently, high profile retractions, survey findings that some management scholars may have engaged in data fabrication and finding falsification, and evidence of statistical errors raise concerns about the trustworthiness of the empirical foundations of management research. In addition, reproducibility, which "…refers to the ability of other researchers to obtain the same results when they reanalyze the same data" (Kepes *et al*, 2014: 456), is not currently required as a condition for publication. The combination of possible reporting problems with a lack of

formal requirements for confirming the accuracy of empirical findings creates conditions for academic misconduct, and dishonest or incorrect study findings could make their way into the literature and serve to compromise the credibility and trustworthiness of our cumulative scientific knowledge. Indeed, more than 20 percent of statistical results in 300 *Strategic Management Journal* articles appear to have been incorrectly reported (Goldfarb & King, 2016), suggesting that strategic management at least, a field within management, does have a reporting and finding problem. Correcting such matters represent crucial steps for protecting the integrity of the field's literatures.

Our article proposes a modest step to help close the gap that allows problematic study findings to find their way into the management literature. We describe and demonstrate three verification procedures that can be used to assess reported statistics in articles and flag errant or fraudulent articles before they become part of the field's knowledge base, hence safeguarding the trustworthiness of our cumulative scientific knowledge. These tests can all be performed using commonly reported data and most statistical software packages. Indeed, the tests are applicable to studies that report the most basic of all statistical tests, can be used to verify findings without requiring original datasets, are objective in nature, and have previously appeared in peer-reviewed research outlets, increasing their face validity. The tests were found to work, as they uncovered numerous reporting anomalies in the L&E article.

Additional tests

Other methods exist for detecting potential problems in empirical research.[5] For example, in the event that the entire dataset can be obtained, simply re-running an author's regression

[5] We thank an anonymous reviewer for these suggestions.

models may not uncover the complete set of possible problems with the underlying data. Abelson (1995) offers procedures for detecting "gaps", "dips", "cliffs" and "peaks" within a set of data which might suggest that some non-random process is affecting the values. Such non-random processes could be the result of data tampering by the researcher, or some unobserved phenomenon which led to the observed values, but in either case they represent violations of normality assumptions and call into question the validity of regression findings based on the data. As noted, these checks are only possible when the full data is available, which is rare in management research.

Abelson (1995) also suggests a number of ways in which a reader or reviewer can get a sense for whether or not reported regression results are credible. This is accomplished by looking for test statistics that are "too large" or "too small", models that fit "too well", or results that seem "too good to be true". There are some rules of thumb to follow, such as being wary of ratios of *F-statistic* to number of observations approaching or exceeding one, but by and large these guidelines rely on the experience and judgment of the observer.

Another technique for detecting potentially problematic empirics is described by Simonsohn (2013). His technique is predicated on the fact that when a given variable is measured across multiple populations we can expect the observed means and standard deviations to be distributed in predictable ways. Too little or too much variance in either the means or the standard deviations across the populations should raise a red flag that either there is an error in the reported data or the authors have doctored the data to fit an agenda. While this is a powerful technique in the realm of experimental studies where a given variable will be observed across multiple different experimental treatments, it is relatively rare in management research to have the same variable measured independently in multiple different populations, and even rarer for

those means and standard deviations to be reported separately. The closest analogue in our field would be studies which conduct analyses of subgroups of a larger population. However, even then the standard practice is to report the descriptive statistics for the entire population rather than for the subgroups individually.

Collectively, all of the tests discussed thus far could play a critical role in protecting and confirming the integrity of empirical findings and the conclusions which are based upon them. We suggest that the verifiability, credibility, and trustworthiness of a study's results should become one of the critical links in a publication process that seems to have emphasized the novelty of ideas -- "what's new" -- rather than "what's true" (Pfeffer, 2007). We join others who suggest that changes in the review process are needed. Indeed, some have recommended several significant revisions to raise the trustworthiness of findings through removing the incentives for misconduct. For example, the use of research registries, changes to the review process to include null, contrarian, and small effect sizes, a halt in a-theoretical model trimming, a multi-part review process whereby the data are collected after the model has been approved by reviewers, replications, and strengthening the methods-emphasis in our communities have each been recommended (see Kepes & McDaniel, 2013, for a review). Our article contributes to these suggestions by adding the role of independent empirical verification tests as a mechanism for assessing the trustworthiness of scientific evidence, during the review process if possible, but after publication if necessary. If the field's credibility depends on evidence that is above reproach (Kepes *et al*. 2014), confirmatory tests become an essential component of the scientific process.

Recommendations for the review process

All stakeholders within management science expect that research studies and their findings are reported as honestly and completely as possible. The field's gatekeepers, the primary participants in the manuscript review process, face a pressing decision: risk publishing problematic studies using a system that does not confirm findings, or take a new path where expanded disclosure and reproducibility tests could detect and reduce incomplete and possibly dishonest reporting. We clearly advocate the latter. We submit that the most effective path forward will involve all parties to the manuscript review process, and that none of those participants will bear an undue burden.

Authors. Authors might appear as independent agents whom simply write articles and offer conclusions. However, their contributions become part of a collective knowledge base that serves a larger community. Through submitting their work for acceptance within this community, the authors have a responsibility to meet the group's expectations and ethical requirements. Since authors are the source of manuscripts, our recommendations on improving the confirmability of study findings and protect the field's trustworthiness begins with them.

Specifically, we recommend that authors provide complete disclosure of their study data consistent with the reporting requirements described by Bettis and his fellow editors (2016: 261) to include coefficient estimates, standard errors, sample sizes and exact *p*-values (no stars or cut-off levels) for all empirical results in analytical models. Further, we call for authors to include variable means, standard deviations, and correlation matrices for all variables included in the analytical models (including interaction terms, transformed variables, etc.), and for all subgroups if appropriate. Second, authors need to describe all data-related decisions pertaining to their variables and analyses, including stating how missing values and outliers were handled, and report the exact sample sizes related to each empirical analytical model. Finally, we suggest that

authors confirm the accuracy of the relationships between empirical tests, tabular reporting of data and findings, hypotheses, and conclusions. Collectively, these suggestions will facilitate re-testing and allow for problems to be corrected before publication and not risk problems afterwards. Ultimately, authors need to attest when submitting their article that their study data are reported fully and that results are accurately and wholly based on those data. Authors should understand that increased disclosure to permit comprehension and evaluation of data may become the new reporting norms.

Journal editors. We call for journal editors to revise the submission process to include new requirements: (1) Following the lead of Bettis and colleagues (2016), editors require all submissions to meet expanded data and finding disclosure requirements regarding coefficients, and also include correlation matrices, sample sizes, discussion of missing values, outliers and the sample sizes for each analytical model. (2) Require that authors attest that their article's data is reported consistent with point (1) and that study findings are based entirely and accurately on those data. (3) Make it clear that by submitting a manuscript for publication consideration, authors accept that their works' findings will be confirmed through re-testing should their articles reach the conditional acceptance stage. (4) Amend manuscript evaluation forms that accompany reviewers' assessments to include a check of whether the data and findings are reported in accordance to the expanded disclosure requirements, and that the data, results and hypotheses appear consistent with one another. And (5) when a manuscript reaches the conditional acceptance point apply the tools described in Test One and Test Three above to verify that the reported findings are accurate.

The costs of implementing recommendations (1) through (4) should be one-time only while those for (5) are relatively minor. Most journals have discretionary budgets for the editor's

travel and support, and such funds might also be used for helping ensure the integrity of the journal's published work by paying for a spot check of empirical findings in conditionally accepted submissions. Further, the verification procedures are not difficult to implement. The *p*-value reconfirmations described in our Test One require only an Excel file and can be done quickly and easily. Once that file is created it would be a simple matter of entering the findings from any particular manuscript to see if they check out. The time and skill required to enter the data from the manuscript and run the analytical models are well within the capabilities of the average graduate student. We submit that these costs are far smaller than those of failing to detect errant or fraudulent results and the subsequent damage to the field's knowledge base. In addition, when Tests One or Three indicate a potential problem with a particular manuscript, we recommend that Test Two be employed using the extant body of published work from the particular authors in an effort to ascertain whether the irregularities are themselves an anomaly or rather an indication of a larger pattern.

Reviewers. Reviewers are the field's experts and offer recommendations to editors on whether a submission should be rejected, revised, or accepted. It therefore seems essential that reviewers carefully assess data and finding reporting within their evaluative process. We call for reviewers to (1) Confirm that a manuscript's data reporting is complete with respect to the expanded data disclosure requirements described above, and also consistent from descriptive statistics to the presentation of the findings in the tables. Reviewers are also requested to ensure that authors disclose decisions about missing values, outliers, and sample sizes for all respective analytical models. (2) Assess that hypotheses are interpreted correctly with respect to the reported findings. These tasks require introductory statistical knowledge only (for example, ensuring that all variables which appear in a regression also appear in the tables of descriptive

statistics, that all coefficients are accompanied with standard errors or *t*-tests and precise *p* values, and that the reported conclusions are interpreted consistent with the empirical results) and should be comfortable for most reviewers of empirical manuscripts.

The additional costs to the reviewers would be minimal; within the process of conducting a review, they would be required only to examine data reporting and interpretation to ensure that all data are fully disclosed and consistent. We are not calling for reviewers to retest data. That particular responsibility can and should be borne at the journal level. Still, if reviewers double-check the reporting requirements, then the editor's ability to retest the data will be ensured and fewer delays will occur with journal editors not having to resend articles back to authors for more data reporting and possible retesting.

Overall, these suggestions add more steps and complexity to the review process. However, these recommendations are less ambitious than proposals in other social science literatures, whereby authors are required to provide their data and analysis codes to journals for independent confirmation (see Dewald *et al*. 1986; Chang & Li, 2015). Indeed, the journal *Management Science* has a "Data Disclosure" policy that now specifies, "[T]o support the scientific process, Management Science, encourages but does not require the disclosure of data associated with the manuscripts we publish…" (http://pubsonline.informs.org/page/mnsc/submission-guidelines). We encourage all gatekeepers to consider this precedent; why should authors of management studies not be required to provide their data and coding, especially in the cases of qualitative or proprietary data sets whereby external replication would be impossible? We recognize that such requirements are not currently the field's generally accepted principles, but those specifications can be easily changed to meet the new publishing environment.

In closing, the current process for manuscript peer review in management research has no formal provision for confirming empirical findings and instead, relies on author integrity to ensure that the findings are reported accurately. Given article retractions, mistakes in empirical findings, and surveys indicating that many scholars have committed "cardinal sins" with their data, it is time that the field takes steps to protect the validity and trustworthiness of its knowledge base. We hope that our article helps spur such remedies.

REFERENCES

Abelson, R.P. 1995. On suspecting fishiness. ***Statistics as Principled Argument.*** Hillsdale, N.J.: L. Erlbaum Associates, pp. 78-88.

Bakker, M., & Wicherts, J. 2011. The (mis)reporting of statistical results in psychology journals. ***Behavior Research Methods***, 43(3): 666–678.

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(4): 715-725.

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

Bettis, R.A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. 2016. Creating repeatable cumulative knowledge in strategic management. ***Strategic Management Journal***, 37: 257-261.

Bohannon, J. 2013. Who's afraid of peer review? ***Science***, 342(6154): 60-65.

Boyd, B. K., Bergh, D. D., & Ketchen Jr, D. J. 2010. Reconsidering the Reputation-Performance Relationship: A Resource-Based View. ***Journal of Management***, 36(3): 588-609.

Chang, A. C., & Li, P. 2015. Is Economics Research Replicable? Sixty Published Articles from Thirteen Journals Say "Usually Not", ***Finance and Economics Discussion Series 2015-083***. Washington: Board of Governors of the Federal Reserve System, http://dx.doi.org/10.17016/FEDS.2015.083.Available at https://www.federalreserve.gov/econresdata/feds/2015/files/2015083pap.pdf

Dewald, W.G., Thursby, J.G., & Anderson, R.G. 1986. Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. ***The American Economic Review***, 76(4): 587-603.

Epskamp, S., & Nuijten, M. B. 2015. ***statcheck: Extract statistics from articles and recompute p values***. R package version 1.0.1. http://CRAN.R-project.org/package=statcheck

Godlee, F., Gale, C. R., & Martyn, C. N. 1998. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: A randomized controlled trial. ***JAMA***, 280(3): 237-240.

Goldfarb, B. D., & King, A. A. (2016). Scientific apophenia in strategic management research: significance tests & mistaken inference. ***Strategic Management Journal***, 37: 167-176.

http://retractionwatch.com/2015/09/29/german-department-head-reprimanded-for-not-catching-mistakes-of-co-author/#more-32793

http://www.uni-mannheim.de/1/presse_uni_medien/pressemitteilungen/2014/Oktober/Prof.%20Dr.%20Ulrich%20Lichtenthaler%20verl%C3%A4sst%20die%20Universit%C3%A4t%20Mannheim/

Hubbard, R., Vetter, D.E., & Little, E.L. 1998. Replication in strategic management: Scientific testing for validity, generalizability and usefulness. ***Strategic Management Journal***, 19: 243-254.

Kepes, S., Bennett, A., & McDaniel, M. (2014). Evidence-based management and the trustworthiness of cumulative scientific knowledge, Implications for teaching, research and practice. ***Academy of Management Learning and Education***, 13: 446-466.

Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? ***Industrial and Organizational Psychology***, 6: 252–268.

Lichtenthaler, U., & Ernst, H. 2012. Integrated knowledge exploitation: The complementarity of product development and technology licensing***. Strategic Management Journal***, 33: 513-534. (Retraction published 2012, ***Strategic Management Journal***, 33: 1341).

Nuijten, M., Hartgerink, C. J., van Assen, M. L. M., Epskamp, S., & Wicherts, J. 2015. The prevalence of statistical reporting errors in psychology (1985–2013). ***Behavior Research Methods***: 1-22.

Pfeffer, J. 2007. Truth's consequences. ***Journal of Organizational Behavior***, 28: 837–839.

Schminke, M. 2009. Editor's comments: the better angels of our nature—ethics and integrity in the publishing process. ***Academy of Management Review***, 34(4): 586-591.

Schroter, S., Black, N., Evans, S., Godlee, F., Osorio, L., & Smith, R. 2008. What errors do peer reviewers detect, and does training improve their ability to detect them? ***Journal of the Royal Society of Medicine***,101(10): 507-514.

Shaver, J. M. 2005. Testing for mediating variables in management research: concerns, implications, and alternative strategies. ***Journal of Management***, 31(3): 330–353.

Simonsohn, U. (2013) Just post it: The lesson from two cases of fabricated data detected by statistics alone. ***Psychological Science***, 24(10), 1875-1888.

Strategic Management Journal, 2012. Retraction: Integrated knowledge exploitation: The complementarity of product development and technology licensing. ***Strategic Management Journal***, 33: 1341.

www.retractionwatch.com/leaderboard

http://retractionwatch.com/2015/09/29/german-department-head-reprimanded-for-not-catching-mistakes-of-co-author/#more-32793

TABLE 1

Advantages and Disadvantages of the Three Tests

Test	Advantage	Disadvantage
Test One: Congruence of reported test statistics Recalculate *p* values based on reported statistics	A direct, straightforward and allows "apples to apples" comparisons of reported significance values for control, independent, moderating and mediating relationships. Can be applied to large samples through using software package such as R package.	Requires a complete disclosure of essential statistics such as *β, se, t,* and *df.* Cannot ascertain whether authors misreported or distorted their statistics in other ways beyond simply misstating how significant particular coefficients are. Vulnerable to the clarity of author reporting. Cannot provide insights into the sizes and directions of the coefficients.
Test Two: Simulation-based verification Estimate how many coefficients may be over- or under-stated relative to an expected "true" effect size	Allow researchers to characterize the stability or generalizability of published findings by answering the question: How likely would we be to get the same results on a different sample from the same population? Allow researchers to detect cherry-picking of samples or models even when the published descriptions of the data and results are perfectly accurate.	A large number of coefficients are required to get meaningful results. Ability to detect errors is limited by the likely nature of the errors or malfeasance. Does not give any specific insight into which particular coefficients may have been misstated or inflated.
Test Three: Re-Verification based on matrices of descriptive statistics Re-run a study's reported regressions using data derived from the published descriptive and correlational statistics	Relatively easy and accessible. Many major statistical software packages have built-in functions to perform the test. Can effectively detect a number of different errors or misstatements.	Need completely reported descriptive statistics for all variables, including interaction terms, transformed variables, or squared terms that are rarely reported. Despite can detect various errors but offer no specificity as to which error(s) and why. Cannot tell whether the published results truly reflect a phenomenon in the underlying population.

TABLE 2

Results of Test One for all coefficients

Model	Number of coefficients in the model	Number of coefficients with recalculated *p*-values different from reported *p*-values	Percent of coefficients with recalculated *p*-values different from reported *p*-values (%)
1	10	1	10
2	12	3	25
3	13	1	8
4	13	1	8
5	13	1	8
6	13	2	15
7	13	1	8
8	14	1	7
9	14	1	7
10	10	4	40
11	12	3	25
12	13	4	31
13	13	2	15
14	13	1	8
15	13	5	38
16	13	2	15
17	14	5	36
18	14	5	36
19	10	4	40
20	12	3	25
21	13	3	23
22	13	4	31
23	13	0	0
24	13	2	15
25	13	4	31
26	14	4	29
27	14	2	14
28	14	4	29
29	14	4	29
Total	**373**	**77**	**21**

TABLE 3

Results of Test One for hypothesis coefficients

Model	Variable	Coefficient	Standard Error	Degrees of Freedom	Calculated *T*	Recalculated *p*-value	Reported *p*-value
3	Prod. dev. X Tech. lic.	0.19	0.10	214	1.90	0.059	<0.1
4	Prod. dev. X Tech. lic.	0.26	0.17	100	1.53	**0.129**	**<0.05**
5	Prod. dev. X Tech. lic.	-0.08	0.16	100	0.50	0.618	>0.1
6	Prod. dev. X Tech. lic.	0.28	0.26	87	1.08	**0.283**	**<0.05**
7	Prod. dev. X Tech. lic.	0.18	0.14	86	1.29	**0.201**	**<0.1**
8	Prod. dev. X Ext. ident.	0.28	0.11	213	2.55	0.012	<0.05
9	Prod. dev. X Ext. comm.	0.03	0.14	213	0.21	0.831	>0.1
12	Prod. dev. X Tech. lic.	0.37	0.31	196	1.19	**0.235**	**<0.05**
13	Prod. dev. X Tech. lic.	0.41	0.28	196	1.46	**0.146**	**<0.05**
14	Prod. dev. X Tech. lic.	0.02	0.84	91	0.02	0.981	>0.1
15	Prod. dev. X Tech. lic.	0.93	0.57	79	1.63	**0.107**	**<0.05**
16	Prod. dev. X Tech. lic.	0.49	0.25	78	1.96	0.054	<0.1
17	Tech.lic. X Int. ident.	0.30	0.39	195	0.77	**0.442**	**<0.05**
18	Tech.lic. X Int. comm.	0.29	0.37	195	0.78	**0.436**	**<0.1**
21	Prod. dev. X Tech. lic.	0.41	0.24	174	1.71	**0.089**	**<0.05**
22	Prod. dev. X Tech. lic.	0.43	0.31	80	1.39	**0.168**	**<0.05**
23	Prod. dev. X Tech. lic.	0.18	0.51	80	0.35	0.725	>0.1
24	Prod. dev. X Tech. lic.	0.34	0.38	71	0.89	**0.376**	**<0.1**
25	Prod. dev. X Tech. lic.	0.29	0.43	69	0.67	**0.505**	**<0.1**
26	Prod. dev. X Ext. ident.	0.47	0.36	173	1.31	**0.192**	**<0.05**
27	Prod. dev. X Ext. comm.	0.17	0.45	173	0.38	0.706	>0.1
28	Tech.lic. X Int. ident.	0.45	0.38	173	1.18	**0.240**	**<0.05**
29	Tech.lic. X Int. comm.	0.32	0.41	173	0.78	**0.436**	**<0.1**

Note: entries in **bold** indicate differences in reported and reproduced values

TABLE 4

Result of Test Three for Six Testable Models

		Reported		Replicated	
Model	Variable	Coefficient	*p*-value	Coefficient	*p*-value
Table 4 Model 1	Firm size	-0.13	<0.05	-0.15	0.005
	R&D intensity	-0.01	>0.1	-0.01	0.691
	Technology exploration	0.41	<0.001	0.42	0.000
	Chemicals/pharmaceuticals	0.14	>0.1	0.16	0.364
	Electronics/semiconductors	0.21	>0.1	0.25	0.208
	Importance cross-licensing	-0.00	>0.1	-0.06	0.247
	Technological diversification	-0.01	>0.1	-0.01	0.835
	Product diversification	0.03	>0.1	0.03	0.616
	International diversification	0.09	>0.1	0.11	0.116
	Patent portfolio strength	**0.14**	**<0.05**	**0.06**	**0.155**
Table 4 Model 2	Firm size	-0.11	<0.05	-0.11	0.018
	R&D intensity	0.01	>0.1	0.01	0.687
	Technology exploration	0.06	>0.1	0.06	0.401
	Chemicals/pharmaceuticals	-0.08	>0.1	-0.07	0.656
	Electronics/semiconductors	0.16	>0.1	0.14	0.413
	Importance cross-licensing	-0.02	>0.1	0.02	0.724
	Technological diversification	0.00	>0.1	-0.00	0.934
	Product diversification	**0.09**	**>0.1**	**0.09**	**0.084**
	International diversification	0.00	>0.1	0	0.979
	Patent portfolio strength	**0.08**	**<0.1**	**0.08**	**0.032**
	Product development	0.63	<0.001	0.72	0.000
	Technology licensing	**0.12**	**<0.1**	**-0.01**	**0.806**
Table 5 Model 10	Firm size	-0.96	<0.001	-0.94	0.001
	R&D intensity	**0.06**	**<0.1**	**0.05**	**0.600**
	Technology exploration	0.29	>0.1	-0.02	0.958
	Chemicals/pharmaceuticals	0.35	>0.1	-0.38	0.693
	Electronics/semiconductors	**1.78**	**<0.05**	**1.30**	**0.228**
	Importance cross-licensing	**0.48**	**<0.05**	**0.11**	**0.680**
	Technological diversification	-0.38	>0.1	-0.09	0.805
	Product diversification	**0.58**	**<0.05**	**0.46**	**0.159**
	International diversification	0.14	>0.1	0.11	0.769
	Patent portfolio strength	0.10	>0.1	0.31	0.17
Table 5 Model 11	Firm size	-0.99	<0.001	-1.05	0.000
	R&D intensity	0.06	>0.1	0.03	0.750
	Technology exploration	0.07	>0.1	-0.61	0.183
	Chemicals/pharmaceuticals	0.11	>0.1	-0.82	0.394
	Electronics/semiconductors	**1.78**	**<0.05**	**1.07**	**0.308**

	Importance cross-licensing	0.29	>0.1	0.28	0.293
	Technological diversification	-0.34	>0.1	-0.04	0.905
	Product diversification	**0.60**	**<0.05**	**0.43**	**0.174**
	International diversification	0.11	>0.1	0.01	0.975
	Patent portfolio strength	0.11	>0.1	0.31	0.162
	Product development	0.26	>0.1	0.73	0.173
	Technology licensing	0.69	<0.05	1.14	0.001
Table 6 Model 19	Firm size	0.11	>0.1	0.14	0.558
	R&D intensity	**0.01**	**>0.1**	**0.13**	**0.083**
	Technology exploration	**0.51**	**<0.1**	**0.300**	**0.340**
	Chemicals/pharmaceuticals	1.24	>0.1	0.98	0.227
	Electronics/semiconductors	0.83	>0.1	0.31	0.734
	Importance cross-licensing	-0.01	>0.1	0.09	0.688
	Technological diversification	-0.61	<0.1	-0.55	0.071
	Product diversification	0.16	>0.1	0.13	0.633
	International diversification	0.53	<0.1	0.50	0.104
	Patent portfolio strength	0.40	<0.05	0.38	0.048
Table 6 Model 20	Firm size	0.08	>0.1	0.14	0.567
	R&D intensity	**0.01**	**>0.1**	**0.13**	**0.076**
	Technology exploration	0.25	>0.1	-0.05	0.891
	Chemicals/pharmaceuticals	0.98	>0.1	0.74	0.370
	Electronics/semiconductors	0.82	>0.1	0.19	0.834
	Importance cross-licensing	-0.19	>0.1	0.17	0.452
	Technological diversification	-0.58	<0.1	-0.53	0.079
	Product diversification	0.19	>0.1	0.16	0.549
	International diversification	0.48	>0.1	0.41	0.189
	Patent portfolio strength	0.41	<0.05	0.39	0.042
	Product development	0.37	>0.1	0.62	0.178
	Technology licensing	**0.63**	**<0.1**	**0.23**	**0.418**

Note: Entries in **bold** indicate differences in reported and reproduced values

FIGURE 1

Results of Test Two