Strategic Organization
2017, Vol. 15(3) 423–436
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1476127017701076
journals.sagepub.com/home/soq
SAGE

Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings

Donald D Bergh
University of Denver, USA

Barton M Sharp
Northern Illinois University, USA

Herman Aguinis
George Washington University, USA

Ming Li
University of Liverpool, UK

Abstract

Recent studies report an inability to replicate previously published research, leading some to suggest that scientific knowledge is facing a credibility crisis. In this essay, we provide evidence on whether strategic management research may itself be vulnerable to these concerns. We conducted a study whereby we attempted to reproduce the empirical findings of 88 articles appearing in the *Strategic Management Journal* using data reported in the articles themselves. About 70% of the studies did not disclose enough data to permit independent tests of reproducibility of their findings. Of those that could be retested, almost one-third reported hypotheses as statistically significant which were no longer so and far more significant results were found to be non-significant in the reproductions than in the opposite direction. Collectively, incomplete reporting practices, disclosure errors, and possible opportunism limit the reproducibility of most studies. Until disclosure standards and requirements change to include more complete reporting and facilitate tests of reproducibility, the strategic management field appears vulnerable to a credibility crisis.

Keywords
knowledge credibility, replication, reproducibility

Corresponding author:
Donald D Bergh, Daniels College of Business, University of Denver, 2101 S. University Blvd, Denver, CO 80208, USA.
Email: dbergh@du.edu

Currently, there are important concerns about the state of study findings in literatures as diverse as psychology, economics, and biomedicine (e.g. Bakker and Wicherts, 2011; Bosco et al., 2016; Chang and Li, 2015). For example, scholars report an inability to replicate the findings of seminal studies (e.g. Begley and Ellis, 2012; Ioannidis, 2012), leading some to suggest a possible credibility crisis may exist in science (e.g. Baker, 2012; Carpenter, 2012; Cortina et al., 2017).

Do these concerns extend to strategic management research? In this essay, we consider whether empirical findings in strategic management research can be reproduced using their own data. Reproducibility is "the ability of other researchers to obtain the same results when they reanalyze the same data" (Kepes et al., 2014: 456). If study findings cannot be reproduced from their own data, questions arise about their credibility. Can the strategic management field's findings be reproduced? If not, why, and if so, how closely do reproduced results compare to those of the original research? In addition, what is the scope of reproducibility—can most empirical studies be reproduced, half, or is it only a few? Are differences between reported and reproduced findings random or do they suggest a systematic bias on the part of authors? Although a recent study examines whether published findings are correctly reported in terms of their coefficient statistical properties (e.g. Goldfarb and King, 2016), we know of no studies that address the congruence of statistical findings and the data upon which they are based in strategic management research. Currently, we lack evidence for estimating the reproducibility of findings and if the field is vulnerable to a credibility crisis.

Our essay reports an initial examination of the reproducibility of findings appearing in strategic management research. Within this field, we examine articles published in the *Strategic Management Journal* (*SMJ*), as this journal offers an unambiguous source of research directly addressing strategic management domains, disseminates the most articles devoted exclusively to strategy research, and has been used in other reviews as a single source for evaluating precedent-setting methodological practices in strategic management studies (e.g. Aguinis et al., 2016; Bergh and Fairbank, 2002; Ferguson and Ketchen, 1999; Shook et al., 2003). We do not isolate the *SMJ* for criticism; rather, we examine some of its articles because it has long been considered a leading strategic management research outlet and its practices are likely to be adopted by others.

Using data from two samples of *SMJ* articles, we attempt to reproduce empirical models and then compare our results with those originally reported. These comparisons provide insights into the proportion of strategic management findings that can be reproduced, whether supported hypotheses remained so, and if there is cause for concern about the state of reproducibility—and by extension the credibility—of empirical findings in strategic management research.

Overall, our study examines some possible drivers behind a potential credibility crisis by documenting the ability to reproduce empirical findings and identifying barriers to reproduction. It illuminates the possible effects of the non-reproducibility of findings on the field's empirical foundation and illustrates how reproducibility (or lack thereof) affects the interpretability and meaning of replication and extension studies. Viewed generally, these results offer insight into whether conditions exist that might create a credibility crisis within the strategic management field.

Method

Reproducing a study's findings generally implies reanalyzing original datasets and comparing reproduced findings with those initially reported in the study. Such a process would seem to require either the data from a focal article or an independent re-collection of the data. However, an alternative approach exists whereby a study's variable means, standard deviations (SDs), correlations, and sample sizes can serve as substitutes for the original data set, as these descriptive and correlational data are all that is necessary for many analytical tests including analysis of variance (ANOVA), linear regression (LR), *t*-tests, discriminant analysis, and structural equation modeling (SEM; see Bergh et al., 2017; Boyd et al., 2010; Shaver, 2005, for illustrations). Indeed, most statistical packages, such as Stata, IBM SPSS, SAS, and R (among others), offer the capability to analyze descriptive and correlational data in lieu of raw data, and the findings from testing either form of data will be identical. For example, the manual for Stata's "corr2data" procedure reports that it …

> allows you to perform analyses from summary statistics (correlations/covariances, means) when these summary statistics are all you know and summary statistics are sufficient to obtain results … the analysis … extracts from the summary statistics you specified, and then makes its calculation based on those statistics … **the results from the regression based on the generated [summary] data are the same as those based on the real data**. (http://www.stata.com/manuals13/dcorr2data.pdf, emphasis added)

Other analytical packages offer a "Matrix" feature which allows users to select between a matrix of descriptive summary and correlational data or the raw data file as the data input when conducting their analyses (demonstrated in the Appendix 2 in the online appendices with IBM SPSS).

Samples

To test whether empirical findings in strategic management research can be reproduced from their own data, we drew samples of articles appearing in *SMJ* that employed LR and SEM analyses—the two most popular techniques in the strategy field (Shook et al., 2003). Both analytical techniques can be implemented using descriptive and correlational statistics and represent a range of analytical sophistication that gives insights into whether reproducing results is more possible in the more accessible basic tests (LR), more advanced ones (SEM), or equally in both.

We assembled two samples of *SMJ* studies using different criteria so our findings might apply to a diverse set of articles. One sample consisted of *SMJ* articles using the LR technique and the other consisted of the top 10 most cited *SMJ* articles that employed the SEM technique, based on citations in 2016 from the Institute for Scientific Information (ISI) Web of Science. The LR study sample represents a typical body of *SMJ* articles—a mix of some having high impact and others having normal and lower impact while the SEM studies likely served more visible roles in shaping the development of the field.

LR studies. We identified the sample of *SMJ* articles reporting findings from LR tests using the following steps. (1) We searched for all *SMJ* articles that used the terms "OLS" or "ordinary least squares" using the full-text search function at Wiley Online (the *SMJ* publisher) through the *SMJ* website (http://smj.strategicmanagement.net). We restricted our search terms to avoid judgment calls and ensure conditions for meaningful replication of our own work. Also, we focused on ordinary least squares (OLS) regression, as opposed to logistic, time series, or other forms because OLS models are the most basic of regression analytical tests and involve little interpretation on our part, thereby minimizing the possibility that we might misunderstand the authors' procedures. (2) We defined our sampling frame to articles published in two time periods, 2000–2004 and 2010–2013, to determine if the reproducibility of findings has changed over time as well as include articles likely to reflect more current methodological practices. (3) We retained only those articles that reported OLS regression models and results. This three-step process identified 79 articles (50 from 2000 to 2004 and 29 from 2010 to 2013). (4) We then examined each to identify the correlations, means, SDs, and sample sizes. The 79 articles are identified in Appendix 1 in the accompanying online appendices.

When reported, the descriptive and correlational statistics were arranged into matrices that were entered as input data into the regression analyses. We then retested each LR study using the corr2data procedure in accordance with the instructions provided within the Stata manual for testing data reported in a matrix format (the program code, illustrative data from one of the 79 *SMJ* LR studies, the reported findings, and the results produced by the Stata analyses are also reported in Appendix 2 in the supplemental online appendices). In the cases where discrepancies existed between reported and reproduced findings, we double-checked the inputted data and re-ran our tests using an alternative analytical software package, IBM SPSS (illustrative syntax is included in Appendix 2 in the supplemental online appendices). This multi-test approach served as a reproduction test of our own data and findings and reduced the likelihood of error on our part.

SEM studies. We identified the sample of *SMJ* articles reporting SEM using the following steps. (1) We searched for all articles that used the terms "structural equation modeling," "structural equations," "lisrel," "amos," or "path analysis" using the *SMJ* website (http://smj.strategicmanagement.net) and its search function. (2) We retained only those that used SEM analyses. (3) We collected citation counts for each article from the ISI Web of Science, ranked them by total counts, and then selected the top 10 (note that one of these 10 articles was also in the sample of 79 OLS articles, leading to a final sample of 88 total articles). We focused on the most highly cited articles because they would have had the largest impact on subsequent research. In addition, this sample provides a different context for comparing reproducibility of study findings since it includes the most influential articles rather than ordinary articles employing the LR study technique. (4) Finally, we examined each identified article to locate the correlations, means, SDs, and sample sizes when reported. The articles are listed in Appendix 1 in the supplemental online appendices.

We conducted the SEM analyses using the IBM SPSS Amos statistical analysis software. The summary data were configured into a matrix, entered into the program, and then tested using the same assumptions and parameter specifications as reported by authors (an example appears in Appendix 2 in the supplemental online appendices). Differences between reported findings and our analyses were retested through repeating each of the analyses 10 times (explained below).

Results

Reproducibility of statistical significance conclusions

In total, 58 of the 79 *SMJ* studies using LR (73%) and 4 of the 10 employing SEM (40%) did not report sufficient information to permit any reproducibility analysis whatsoever. The 58 LR studies that could not be retested (36 from 2000 to 2004 and 22 from 2010 to 2013) had incomplete disclosure of means, SDs, and correlations for some study variables, missing cells within the correlation matrices, statistical analyses using disaggregated subgroups, transformed variables for which summary statistics were not reported, or correlation matrices which were not positive semi-definite (please see Table 1 for the full list of reasons).[1] These barriers to reproduction are relatively equal across both time periods suggesting data reporting practices have consistently impeded reproduction in a large majority of LR studies and nearly half of those using SEM analyses.

Of the 21 LR studies that could be retested, 16 could be reproduced partially (some models could be reproduced but others could not) and 5 fully (all models could be reproduced). In total, 4 of the 10 SEM studies could not be reanalyzed due to one missing a correlation matrix, another failing to report means and SDs for some study variables, one that did not specify the precise model that was tested, and one that did not include the coefficients of the final model. Six provided sufficient descriptive summary data to permit reanalysis. Collectively, we were able to retest 20 *SMJ* studies employing only LR, five that used only SEM, and one study that reported both techniques. Overall, we could either partially or fully retest the results reported in only 26 of 88 articles (29.5%).

Table 1. Results of reproducibility study findings that were based on ordinary least squares regression.

	Number of articles (% of sample)	
Reason regression results could not be reproduced	2000–2004	2011–2013
Descriptive statistics not reported	10 (20%)	5 (17%)
Descriptive statistics missing for key variables	7 (14%)	1 (3%)
Industry, firm, or time dummies not reported in descriptive statistics or regression results	5 (10%)	13 (45%)
Descriptive statistics given for raw variables, transformed variables used in regressions	5 (10%)	2 (7%)
Reported tables of correlations not positive semi-definite	4 (8%)	1 (3%)
Descriptive statistics given for full sample, regressions conducted on sub-samples	3 (6%)	
Other	2 (4%)	
Total articles for which reproducibility analysis was not possible	36 (72%)	22 (76%)
Problems in reproduced studies that limited complete retesting		
Descriptive statistics not reported for interaction terms	5 (10%)	5 (17%)
Descriptive statistics missing for other key variables	3 (6%)	1 (3%)
Descriptive statistics given for raw variables, transformed variables used in regressions	1 (2%)	
Reported tables of correlations not positive semi-definite	1 (2%)	
Total articles for which some but not all models could be reproduced	10 (20%)	6 (21%)

Following previous tests of the congruence between reported and reproduced findings in other scientific fields (e.g. Bakker and Wicherts, 2011; Berle and Starcevic, 2007; Garcia-Berthou and Alcaraz, 2004), we focused on the signs of the coefficients (indicating either positive or negative relations with the dependent variable) rather than their actual magnitude to avoid errors on our part. Further some authors report coefficients in standardized formats, others report them as non-standardized, and it is not always clear which approach was used. For example, some articles include a table with regression coefficients that excludes the intercept, giving the impression that coefficients are standardized—because in those cases the value for the intercept is zero (Aguinis, 2004). But, upon reading the article, it becomes clear that coefficients are actually unstandardized and the table simply excluded information regarding the intercept. Furthermore, the directionality of the coefficients and their significance levels always play a role in hypothesis testing, whereas the magnitude rarely if ever does.

The 21 articles using LR reported a total of 732 coefficients in the models which could be reproduced. The reproduced findings corroborated 670 of the 732 (91.5%) coefficient directional signs (either both coefficients were the same sign, or one of them was zero). The six studies employing SEM reported 10 models that collectively included 91 coefficients, of which 86 directional signs were reproduced (95%).

Next, we compared reported with reproduced statistical significance threshold levels. Our retests focused on significance bands ($p<0.01$; $p<0.05$), rather than precise point estimates for the p values because most prior *SMJ* articles indicate statistical significance using the star notation (e.g. ** for $p<0.01$; * for $p<0.05$). The re-analyses of the 21 LR studies resulted in the reproduction of the precise reported significance band for 538 of the 732 coefficients (73.5%). The re-analyses of the six SEM studies resulted in 79 of the 91 coefficients' significance levels (87%) within the same bands.

Reproducibility of results and conclusions regarding hypothesized relations

Hypotheses serve as the basis for article conclusions and are critical in terms of a study's "take-away" message, including implications for theory and practice. If the findings pertaining to the hypotheses cannot be reproduced, then causes for concern regarding the credibility of findings and recommendations would exist. Accordingly, we conducted additional analyses focused on the variables corresponding to hypotheses.

We first identified the coefficients in the 21 re-analyzable *SMJ* studies using LR that were associated with hypothesis testing. We found 144 coefficients representing 51 variables associated with hypotheses. Of those 144 coefficients, 14 associated with 10 variables were reported as statistically significant in the original *SMJ* article but were reproduced as statistically non-significant (14 of 144 coefficients, or 10%; 10 of 51 variables, or 20%). This change in statistical support occurred in 6 of the 21 articles (28% of articles). We also found that three coefficients were reported as statistically non-significant in the original *SMJ* article but reproduced as having significance at the $p < 0.05$ level or higher (3 of 144 coefficients, or 2%).

We then considered a possible alternative explanation for these findings, namely that a lack of reproducibility was due to rounding. In other words, authors who found an exact p value of 0.0499 for a given coefficient might report that finding with a single star indicating $p < 0.05$. In addition, it is possible that rounding in the reported descriptive statistics might lead our reproduced findings to vary slightly from the original regressions, possibly leading to an exact p value of 0.051 which we would have to categorize in the $p < 0.10$ band. Thus, we tested for the possibility that reported and reproduced p values may have been concentrated very near the $p < 0.05$ threshold leading us to perceive a lack of reproducibility when in fact no meaningful difference existed.

The results are reported in Table 2 for the hypotheses in the six *SMJ* studies using LR that lost statistical support. Results from independent Stata and IBM SPSS analyses are also identical, suggesting that there is no effect on results due to the choice of software package. The table shows that only 1 of the 14 coefficients was close to the threshold; in Study 5, the reported level was $p < 0.05$ whereas the reproduced value was $p = 0.053$. None of the other reproduced coefficients was even border-line to conventional significance levels. We also note that four of the six *SMJ* studies using LR each have one result that was not reproduced, one of the six has two results that could not be reproduced, and one study has eight reportedly significant findings that we could not reproduce.

Overall, the findings indicate that in six OLS studies, hypotheses which were supported in the publication lost empirical support in our retests. Based on our reanalysis, some of these studies' substantive conclusions would not have received supporting evidence.

We next identified the coefficients in the six SEM studies that were associated with hypotheses. Hypotheses in two studies (33%) could not be reproduced, one study lost support for 1 of 14 supported hypotheses while another lost support for 11 of 19. In addition, we examined whether the differences in significance levels between reported and reproduced are due to rounding error. The findings reported in Table 3 suggest that such an alternative explanation is unlikely.[2] Furthermore, all of the significant-turned-non-significant findings pertained to hypotheses. Thus, 22% of the statistical significance of the hypothesis testing coefficients (12 of 55) in the *SMJ* studies using SEM were not confirmed and no cases existed where the findings for hypotheses were reported as non-significant but found to be significant in the reanalysis.

Averaging across the LR and SEM samples, about 30% re-analyzable studies (8 of 27 total) reported significant hypotheses that lost statistical support in the reproduction tests. By contrast,

Table 2. Reproducibility of ordinary linear regression hypothesis findings: reported and reproduced statistical significance levels.

Study identifier	Reported *p* value	Reproduced *p* value by Stata	Reproduced *p* value by IBM SPSS
1	<0.01	0.087	0.087
2	<0.05	0.704	0.704
3	<0.01	0.244	0.244
4	<0.01	0.109	0.109
4	<0.001	0.179	0.179
4	<0.05	0.241	0.241
4	<0.05	0.386	0.386
4	<0.001	0.172	0.172
4	<0.01	0.093	0.093
4	<0.001	0.115	0.115
4	<0.05	0.909	0.909
5	<0.05	0.053	0.053
6	<0.05	0.174	0.174
6	<0.05	0.213	0.213

p, observed probability for the null hypothesis that the coefficient is zero in the population. Reported *p* values are those reported in the published studies and reproduced results are those obtained using the reproducibility procedures described in text.

Table 3. Reproducibility of structural equation modeling hypothesis findings: reported and reproduced statistical significance levels.

Study identifier	Reported *p* value	Reproduced *p* value
1	<0.05	0.067
2	<0.05	0.211
2	<0.05	0.749
2	<0.05	0.511
2	<0.05	0.763
2	<0.05	0.505
2	<0.05	0.912
2	<0.05	0.898
2	<0.05	0.912
2	<0.05	0.822
2	<0.05	0.822
2	<0.05	0.053

Note: p, observed probability for the null hypothesis that the coefficient is zero in the population. Reported *p* values are those reported in the published studies and reproduced results are those obtained using the reproducibility procedures described in text.

less than 5% have hypotheses reported as statistically non-significant but meet conventional statistical levels in the retests.

Discussion

A high profile debate on the replicability of study findings is emerging across multiple disciplines (e.g. Baker, 2012; Bissell, 2013; Bosco et al., 2016; Carpenter, 2012) raising concerns about whether science is facing a credibility crisis (Butler et al., 2017; Schwab and Starbuck, 2017). Our essay considers this topic within strategic management research. We sought to document the reproducibility of study findings, identify barriers that impede reproducibility, and when possible, compare whether reported and reproduced findings were the same. Drawing from two samples of articles appearing in the *SMJ*, we were unable to conduct reproducibility analysis for more than 70% of studies employing LR (consistent over the two time periods) and for four of the top 10 cited articles using SEM. Furthermore, of the studies that could be reproduced, nearly one of three reported hypotheses as statistically significant which were no longer so in retesting, and far more significant results were found to be non-significant in the reproductions than in the opposite direction. In some cases, multiple hypotheses within a single article lost support even though most of the corresponding coefficient directional signs were reproduced. Primary conclusions in those articles were based on statistical results that could not be reproduced. These findings exist for articles having low as well as high impact.

Overall, based on our sample of 88 *SMJ* articles, the strategic management literature appears vulnerable to credibility problems for two main reasons. One, the majority of the articles did not report their data sufficiently to permit reproduction, leaving us in the dark with regards to the accuracy of their reported results. Two, among those articles where reproduction analyses were possible, a significant number of discrepancies existed between reported and reproduced significance levels.

Implications for substantive conclusions and research evaluation practices

The study findings suggest some initial insights into whether findings in strategic management empirical research can be confirmed from their own reported data. For the most part, we simply cannot tell, although some indications point to "no." In those cases where we were able to re-run the authors' analytical models, we found a troubling proportion of discrepancies between the reported and reproduced results; in some cases, those discrepancies could alter our conclusions about hypothesized relations and their underlying theories. In the majority of cases using LR and nearly the majority of those using SEM, reproduction was simply not possible, rendering us unable to confirm findings and conclusions. At best, this lack of reproducibility represents deficient reporting practices. At worst, it means that we as researchers are attempting to build on results that do not accurately and fully represent the underlying data, and that we as teachers are passing on to practitioners conclusions that may not be sound reflections of the true underlying phenomena. To bolster the value of and confidence in empirical research, we call for the field to recognize the role of reproducibility in the scientific process.

In addition, our findings suggest that replication studies may have challenges: if we cannot retest 70% or more of the studies' findings, and some 30% of the articles that could be retested contained significant coefficients used to test hypotheses that could not be confirmed, then we have a potentially perilous basis for offering conclusions about the meaning of findings from replications. Indeed, authors of replications that have different conclusions might attempt to attribute them to the generalizability of the focal research while all along the reasons for the discrepancies could be unknowable undisclosed data decisions, errors, or possibly even malfeasance. A case in point is an article in the 2016 *SMJ* Special Issue on Replication in Strategic Management by Park, Borah, and Kotha. These authors attempt to replicate three articles on signaling theory, finding no support for original results, concluding that the reasons for the differences in their replicated results included sampling periods, measures, geographical context, extraneous factors, and omitted variables (Park et al., 2016).

However, we posit that replications need to first test whether the focal study is reproducible from its own data. If reproducibility is unsuccessful or not possible, the ability to draw conclusions from a replication could be compromised, as any differences between the findings of an original study and a replication could be due to unobservable issues. For example, the authors of the original study may not have disclosed their decisions about outliers (Aguinis et al., 2013), cherry-picked their findings from a larger set of models (Bosco et al., 2016), tweaked and altered the analyses (Banks et al., 2016), or state that they test one type of effect such as full mediation whereas in actuality they do not. In addition, articles may experience a metamorphosis during the review process whereby authors may engage in post hoc alterations of hypotheses and data as well as engage in questionable research practices (Bosco et al., 2016; Butler et al., 2017; O'Boyle et al., 2014). If such decisions are employed, then descriptive and correlational data may not reflect the data that are ultimately used for testing the models. A reproduction test would uncover these incongruencies while a replication would not. In the case of a replication study, the author would likely produce a study that would mirror the decisions that produced the reported descriptive and correlational data of the focal study although might be unable to replicate the findings due to the unobservable and unreported actions that led to the final results. A reproduction test could detect this possible problem before the replication was even attempted.

It is critical that authors of replications first reproduce the focal study's findings from its own data. If the reproduction yields findings that differ from those reported in the original studies, replication researchers are faced with the conundrum of deciphering whether differences in a replication are due to context or because the results in the original study do not reflect its own underlying data. Thus, we recommend that all replications first reproduce focal study findings. If replication studies do not reproduce focal study findings, and they are unable to replicate their results, we may not know why the observed differences exist.

Thus, the reproducibility of a focal study's findings is a vital and essential prelude to meeting the conditions for repeatable cumulative knowledge development. Until a study's findings are reproduced, they cannot be assumed as a reliable benchmark for replication, comparison, and extension. In those instances where reproduction is not possible, we recommend that authors adopt a cautious interpretation of differences in their results, emphasizing more on what they know (their study decisions), less on what they do not know (the decisions in the focal study), and calling for additional research to reconcile and understand differences in conclusions. We propose that authors employ a "verify then trust" approach before attempting any type of replication.

The findings also have implications for the field's peer review process. We posit that the presence of non-reproducible results limits the confirmation of received results and that the peer review process needs to remove the obstacles that stand in the way of reproduction tests. We suggest the following remedies would increase the reproducibility of findings: (1) reporting significance levels using precise *p* values rather than cutoffs such as 0.01 and 0.05; (2) disclosing all methodological decisions that have a bearing on the reported findings including the handling of missing data and outliers (perhaps in a supplemental online file); (3) reporting all descriptive, correlational, and corresponding sample size data for *all* variables in each corresponding test, including controls, dependent, independent, mediators and moderators, and transformed variables; (4) reporting verification of linear models as a preliminary stage in all models, including those that also test more complex structures; (5) including figures that specify all of the variables (indicators, latent factors, controls, and error terms); (6) in the case of path models, indicating which covariances among exogenous variables were allowed to vary (alternatively, a footnote to the figure should clarify which covariances were allowed to vary, and which were fixed to zero); and (7) reproducibility tests, or at least evidence that reproducibility can be achieved through the reported data. Such disclosures would close a disclosure loophole in current peer review processes, permit independent reproduction, allow for identification of reporting mistakes, discourage some forms of questionable research practices and scientific misconduct, strengthen the field's empirical literatures and contribute to replication and meta-analytical research (e.g. Aytug et al., 2012). Furthermore, these recommendations go beyond recent changes in submission requirements at journals such as the *SMJ* (Bettis et al., 2016), *Journal of Management* (Wright, 2016), and others, but would more likely produce the kind of disclosures that these editors expect for their published articles.

Potential limitations and suggestions for the future research

Our analyses and results need to be considered relative to our study's limitations. First, we retested studies using linear modeling techniques which, although the most popular, do not represent the universe of approaches used in strategic management research. We cannot make any conclusions regarding the reproducibility of empirical findings of studies that did not use LR or SEM. However, we have no reason to expect that those other approaches would be any more reproducible than the studies that we examined. Indeed, most members of the strategic management research community are trained in the general linear model, so if problems are found in the application of this most basic model, then it would seem likely that problems could exist elsewhere. Thus, if we cannot reproduce most studies using techniques that are diffused widely, then studies employing more complex approaches may be even less amenable to reproduction and our study's results may actually understate the verification problems within the management field.

Second, our findings could be influenced by author reporting. Authors are required to review their findings before their articles are published, check proofs, are held responsible for their study's disclosures, and generally make few reporting errors (Bakker and Wicherts, 2011). But, to the extent that authors make mistakes or behave opportunistically with their reporting, then reproduction findings could be affected by such decisions (Schwab and Starbuck, 2017). However, given that (1) most coefficient signs were reproduced and (2) the discrepancies in the retested results greatly favored the support of hypotheses, then the presence of errors may be less likely than the "chase of statistical significance … and the strong tendency to obtain confirmed hypotheses" (Kepes and McDaniel, 2013: 254). So author error alone may not be driving the observed differences in our findings. Although intentional opportunism is a potential cause, it would be premature to make such a claim based on our results. Such evidence instead underscores the need for reproduction in the review process and further research into the articles whose results could not be confirmed using their own reported data.

Finally, we examined articles appearing in one journal only, the *SMJ*, a high-quality outlet devoted to the field of strategic management studies, which has among the highest review standards. However, like any study, our findings may not apply to other journals. We hope that the future research will extend our reproducibility approach to an examination of other journals in strategic management research and other management subfields such as organizational behavior, entrepreneurship, human resource management, and international business. Such extensions will provide insights into the scope and scale of the reproducibility challenges in empirical research.

Concluding remarks

We suggest that limitations in studies' reproducibility present a threat to the credibility of some study findings within the strategic management literature. If we cannot reproduce a study using its own data, then can we have confidence in its findings?

Our findings represent a first assessment of the congruence between findings and the data upon which they are based and more research into the scope and scale of reproducibility is needed to fully comprehend the size of the matter and its implications. We therefore call for extensions of our research methodology to include seminal studies that shaped the field, studies published in different time periods, those using other analytical approaches (the data matrix approach can be used to test other forms of regression and linear and non-linear models such as factor analysis), and broadening the assessments to different literatures beyond strategic management. A comprehensive assessment of reproducibility is needed to fully comprehend its pervasiveness and effects.

Overall, the credibility of the strategic management field's body of knowledge seems at risk until disclosure and peer review requirements are changed to increase the reproducibility of all empirical studies. We hope that our study's findings motivate our fellow researchers to further assess empirical work, to educate all scholars about basic reporting requirements for enhancing reproducibility, and that the field's gatekeepers revise disclosure requirements and include reproducibility in the review process going forward.

Acknowledgement

We gratefully acknowledge the helpful comments from Myles Shaver on a previous version of this manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. When a matrix is not positive semi-definite, then a mismatch likely exists within the relations and some of the values (i.e. bivariate correlations) are likely to be out of range. For example, if variables A and B are positively and highly correlated, and B and C are also highly and positively correlated, then the correlation between A and C must be high and positive as well (Aguinis and Whitehead, 1997). But, for example, if the correlation between A and C is zero, then the statistical properties of the matrix are not positive semi-definite. Possible sources of a non-positive semi-definite matrix are mistakes in the published correlation table, rounding errors that create the appearance of negative variance, and pairwise correlations within the same matrix that have different sample sizes.
2. We retested all SEM studies where differences existed between reported and reproduced conclusions as undisclosed reporting decisions with respect to estimator approaches, correlations between error terms, and model fitting procedures might lead to differences in results (Landis et al., 2009). One possible source for such differences is that the maximum likelihood estimator is converging on a local maximum which might have occurred in either the *Strategic Management Journal* (SMJ) article or in our reanalysis, an outcome that we attempted to address through conducting multiple analyses. Similar to Shaver (2005), we used maximum likelihood estimation (MLE). MLE attempts to maximize the likelihood function—the function that represents the likelihood of the data that are observed. However, depending on the initial parameters chosen, the algorithm might stop and return premature estimates, which are called "local maxima." There are no known solutions for this problem, but one way to address it is to use a range of randomly generated initial parameters (Myung, 2003), which is the procedure implemented by IBM SPSS Amos, LISREL, and many other programs typically used by strategic management researchers. Nevertheless, even if random parameters are used, there may be a difference between results reported in published articles and those reproduced in our study, which may be a reason for the observed discrepancies between results. To assess the possibility that local maxima may have occurred in our reproductions, we re-ran each of our analyses 10 times. As expected, and because the initial parameter values are random, substantive results remained unchanged. So, there is a possibility that initial

parameter estimates in the published studies were not random, which may have led to local maxima, and authors did not disclose this information. The failure to report such information inhibits verification and replication of the findings.

References

Aguinis H (2004) *Regression Analysis for Categorical Moderators*. New York, NY: Guilford Press.

Aguinis H, Edwards J and Bradley K (2016) Improving our understanding of moderation and mediation in strategic management research. *Organizational Research Methods*. Epub ahead of print 27 January. DOI: 10.1177/1094428115562749.

Aguinis H and Whitehead R (1997) Sampling variance in the correlation coefficient under indirect range restriction, implications for validity generalization. *Journal of Applied Psychology* 82: 528–538.

Aguinis H, Gottfredson RK and Joo H (2013) Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods* 16: 270–301.

Aytug Z, Rothstein H, Zhou Z and Kern M (2012) Revealed or concealed? Transparency of procedures, decisions, and judgments calls in meta-analyses. *Organizational Research Methods* 15: 103–133.

Baker R (2012) Independent labs to verify high-profile articles. *Nature*. DOI: 10.1038.2012.1/176.

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

Banks GC, O'Boyle EH, Pollack JM, White CD and Batchelor JH (2016) Questions about questionable research practices in the field of management, A guest commentary. *Journal of Management* 42: 5–20.

Begley CG and Ellis LM (2012) Drug development, raise standards for pre-clinical cancer research. *Nature* 483: 531–532.

Bergh DD and Fairbank JF (2002) Measuring and testing change in strategic management research. *Strategic Management Journal* 23: 359–366.

Bergh DD, Sharp B and Li M (2017) Tests for identifying "Red Flags" in empirical findings: Demonstration and recommendations for authors, reviewers, and editors. *The Academy of Management Learning and Education* 16: 110–124.

Berle D and Starcevic V (2007) Inconsistencies between reported test statistics and *p*-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research* 16: 202–207.

Bettis RA, Ethiraj S, Gambardella A, et al. (2016) Creating repeatable cumulative knowledge in strategic management. *Strategic Management Journal* 37: 257–261.

Bissell M (2013) Reproducibility, the risks of the replication drive. *Nature* 503: 333–334.

Bosco FA, Aguinis H, Field JG, et al. (2016) HARKing's threat to organizational research, evidence from primary and meta-analytic sources. *Personnel Psychology* 69: 709–750. DOI: 10.1111/peps.12111.

Boyd BK, Bergh DD and Ketchen DJ Jr (2010) Reconsidering the reputation-performance relationship, a resource-based view. *Journal of Management* 36: 588–609.

Butler N, Delaney H and Spoelstra S (2017) The grey zone: Questionable research practices in the business school. *The Academy of Management Learning and Education* 16: 94–109.

Carpenter S (2012) Psychology's bold initiative. *Science* 335: 1558–1561.

Chang AC and Li P (2015) *Is economics research replicable? Sixty published articles from thirteen journals say "usually not." Finance and economics discussion series 2015-083*. Washington, DC: Board of Governors of the Federal Reserve System. Available at: http://dx.doi.org/10.17016/FEDS.2015.083; https://www.federalreserve.gov/econresdata/feds/2015/files/2015083pap.pdf

Cortina JM, Aguinis H and DeShon RP (2017) Twilight of dawn or of evening? *A Century of Research Methods in the Journal of Psychology* 102: 274–290.

Ferguson TD and Ketchen DJ Jr (1999) Organizational configuration and performance: The role of statistical power in extant research. *Strategic Management Journal* 20: 385–396.

Garcia-Berthou E and Alcaraz C (2004) Incongruence between test statistics and *p* values in medical articles. *BMC Medical Research Methodology* 4: 13.

Goldfarb BD and King AA (2016) Scientific apophenia in strategic management research: Significance tests and mistaken inference. *Strategic Management Journal* 37: 167–176.

Ioannidis JPA (2012) Why science is not necessarily self-correcting. *Perspectives on Psychological Science* 7: 645–654.

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

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

Landis RS, Edwards BD and Cortina JM (2009) On the practice of allowing correlated residuals among indicators in structural equation models. In: Lance CE and Vandenberg RJ (eds) *Statistical and Methodological Myths and Urban Legends*. New York: Routledge, pp. 193–214.

Myung IJ (2003) Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* 47: 90–100.

O'Boyle EH, Banks GC and Gonzalez-Mulé E (2014) The Chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management* 43: 376–399.

Park UD, Borah A and Kotha S (2016) Signaling revisited: The use of signals in the market for IPOs. *Strategic Management Journal* 37: 2362–2377.

Schwab A and Starbuck W (2017) A call for openness in research reporting: How to turn covert practices into helpful tools. *The Academy of Management Learning and Education* 16: 125–141.

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

Shook CL, Ketchen DJ, Cycyota CS, et al. (2003) Data analytic trends and training in strategic management. *Strategic Management Journal* 24: 1231–1237.

Wright PM (2016) Ensuring research integrity: An editor's perspective. *Journal of Management* 42: 1037–1043.

Author biographies

Donald D Bergh is the Louis D. Beaumont Chair of Business Administration and Professor of Management at the University of Denver. He received a PhD from the University of Colorado at Boulder. His research in corporate strategy and research methods has appeared in the *Academy of Management Journal*, *Strategic Management Journal*, *Organization Science*, *Journal of Management*, *Journal of Management Studies* and *Organizational Research Methods*. In addition to serving on the editorial review boards of these journals, Dr. Bergh has been an associate editor of the *Academy of Management Journal*, *Organizational Research Methods*, *Journal of Management Studies* and, in July 2017, will become Chair of the Scientific Integrity and Rigor Task Force of the *Journal of Management*.

Barton M Sharp is the Mike and Kristina McGrath Professor of Entrepreneurship in the Department of Management at Northern Illinois University. He received a PhD from Purdue University. His research interests include the precedents and consequences of innovation within established organizations, corporate strategy, and research methodologies. He has published work in the *Journal of Management*, *Academy of Management Learning & Education*, and *Organizational Research Methods*, among others.

Herman Aguinis is the Avram Tucker Distinguished Scholar and Professor of Management in the School of Business, George Washington University. His research interests span human resource management, organizational behavior, and research methods and analysis topics. He published five books and about 140 articles in refereed journals, delivered about 240 presentations and keynote addresses at professional conferences, and delivered more than 120 invited presentations in all seven continents except for Antarctica. He is a Fellow of the Academy of Management, American Psychological Association, Association for Psychological Science, and Society for Industrial and Organizational Psychology. He served as editor-in-chief of *Organizational Research Methods* and received numerous awards including the Academy of Management Research Methods Division Distinguished Career Award for lifetime scientific contributions, Academy of Management Entrepreneurship Division IDEA Thought Leader Award, and Best Article of Year Awards from *Personnel Psychology*, *Journal of Organizational Behavior*, *Academy of Management Perspectives*, and *Organizational Research Methods*.

Ming Li is a Senior Lecturer at University of Liverpool Management School. She received a PhD in management from University College Dublin. Her current research interests include cross-cultural management, international HRM, global leadership, Chinese management, and application of research methods. She has published work in journals such as the *Academy of Management Learning & Education*, *Personality and Individual Differences*, and *Organizational Research Methods*.