Strategic Management Journal
Strat. Mgmt. J., **37**: 257–261 (2016)
Published online EarlyView in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/smj.2477
Received 9 November 2015; Final revision received 9 November 2015

EDITORIAL

CREATING REPEATABLE CUMULATIVE KNOWLEDGE IN STRATEGIC MANAGEMENT: A CALL FOR A BROAD AND DEEP CONVERSATION AMONG AUTHORS, REFEREES, AND EDITORS

RICHARD A. BETTIS,[1]* SENDIL ETHIRAJ,[2] ALFONSO GAMBARDELLA,[3] CONSTANCE HELFAT,[4] and WILL MITCHELL[5,6]
[1] *Strategy and Entrepreneurship Department, Kenan-Flagler Business School, University of North Carolina, Chapel Hill, North Carolina, U.S.A.*
[2] *Strategy and Entrepreneurship, London Business School, London, U.K.*
[3] *Department of Management and Technology, CRIOS, Bocconi University, Milan, Italy*
[4] *Tuck School of Business, Dartmouth College, Hanover, New Hampshire, U.S.A.*
[5] *Management Department, Rotman Business School, University of Toronto, Toronto, Ontario, Canada*
[6] *Strategy Area, Fuqua School of Business, Duke University, Durham, North Carolina, U.S.A.*

INTRODUCTION

The objective of this article is to promote discussions and educational efforts among Ph.D. students, scholars, referees, and editors in strategic management regarding the repeatability and cumulativeness of our statistical research knowledge. We believe such discussions and educational efforts will help all of us recognize the *need to develop more appropriate knowledge and norms around the use and interpretation of statistics*. The editorial process at journals can help in this regard, and we are making some policy changes at *SMJ* as discussed below. However, we do not believe that editors should become the enforcers of a substantial set of rigid rules. Instead, as scholars, the practice of our profession should reflect strong norms of appropriate methodological knowledge and practice.

Since statistical methods became common to many fields in the late 1950s, serious questions have been raised by statisticians and scientists about the way statistics has been practiced as a scientific methodology, and the reliability of the statistical models and tests reported in journal publication. In fact, the two original approaches to testing statistical hypotheses—by Fisher and the joint work of Neyman and Pearson—differed from the approach that was adopted by early textbook writers and that has become the dominant practice today (Gigerenzer, 2004). This current approach, known as Null Hypothesis Statistical Tests (NHSTs), was strongly opposed by Fisher, Neyman, Pearson, and others. Nevertheless, various critiques by many scholars over the years have had a hard time gaining traction until recently. Recent research now provides striking evidence regarding the repeatability of statistical research and has called into question many issues related to NHSTs and the institutional context that supports it.

*Correspondence to: Richard A. Bettis, Campus box 3490, McColl Building, Chapel Hill, NC 27599-3490, U.S.A. E-mail: r_bettis@unc.edu

REPEATABILITY IN BIOMEDICAL SCIENCE AND PSYCHOLOGY

The current challenges to the use and interpretation of statistics have arisen largely from recent attempts to replicate important statistical results in biomedical science and psychology. Thoughtful scholars in these two fields are pushing hard for reforms. The relevant literature regarding repeatability in both fields is large, and we only review a small part of it here.

In a seminal replication study of biomedical science published in the *Journal of the American Medical Association*, Ioannidis (2005) examined a sample of highly cited papers regarding medical interventions from top medical journals beginning in 1990 and ending in 2003. He then examined other papers (including some meta-analyses when appropriate) that tested the same interventions as the original highly cited papers. His analyses and conclusions were detailed, complex, and nuanced. Overall, he concluded that "A third of the most-cited clinical research seems to have replication problems, and this seems to be as large, if not larger than the vast majority of less-cited clinical research" (2005: 224). He also found that the "replication" studies that affirmed the original results tended to find weaker effects. Since this first study was published, there have been further supporting studies and considerable ferment in the biomedical community regarding repeatability. Recently, the Editor-in-Chief of *The Lancet,* one of the top medical journals in the world, commented (Horton, 2015) that " … much of the scientific literature, perhaps half may simply be untrue." It should be noted that sources of error in biomedical science do not stem only from statistics, but also can include complex laboratory protocols, including correct isolation and preparation of materials such as tissue samples. Such sources of error are generally not relevant in statistical research reported in strategic management.

In the field of psychology, there has been a long history of criticism of statistical practice. Gigerenzer (2004) briefly reviews some early critiques of statistical practice and then discusses some specifics. Simmons, Nelson, and Simonsohn (2011) discuss and document the manipulation of research to obtain statistically significant results in psychology experiments using the techniques of *p*-hacking. Perhaps the most relevant study in psychology reported on 100 replications from three important psychology journals in 2008 (Open Science Collaboration, 2015). Here, the cooperation of the authors of the replicated papers was secured, and their original materials and protocols were used by other scholars working in other laboratories to replicate the original results. In summary, 97 percent of the original studies had statistically significant results, while only 36 percent of the replications did. Furthermore, only 47 percent of the original effect sizes were within the 95 percent confidence interval of the replication. In strategy research, Goldfarb and King (2016) highlight similar issues.

THE NATURE OF THE PROBLEM IN STRATEGIC MANAGEMENT

In order to illustrate the nature of the statistical and institutional problems raised by the need for repeatability, we construct a purely hypothetical "thought experiment." Three different but hard-working junior scholars in strategic management unbeknownst to each other set out to examine a particularly "interesting" hypothesis using a statistical model. All three determine that the same appropriate specification will be used. Each gathers a large but different sample of similar data for estimating this model.

Two of the researchers finish the statistical work and find that the appropriate model coefficient is not statistically significant at less than the five percent level ($p < 0.05$). Since they cannot publish a "non-result" in a journal, each terminates the project and frustratingly moves on to a different study. Sadly, two years later, both are denied tenure for lack of one more top journal article.

At about the same time as the first two scholars are denied tenure, the third scholar finds that the appropriate coefficient in her study is statistically significant at the 0.01 level. She is especially excited about the level of significance since this indicates a strong result. She hurriedly writes up a preliminary research paper for conference submission. Subsequently, it wins the best paper award at the SMS Annual Meeting. For this, her dean recognizes her at a faculty meeting and grants her a substantial salary increase. Subsequently, after two relatively small revisions, the paper is accepted at *SMJ*. The relevant hypothesis is now considered by many scholars to be established as "proven," and as such, replications with different but comparable samples will not be published. Shortly after the paper is published, she is granted tenure, having obtained the required number of paper accepted by top journals.

There are several problems with this hypothetical outcome. Unwarranted faith is placed in the *p*-value as a breakpoint (0.05) between valuable and worthless results. The coefficient effect size is seemingly ignored. Furthermore, there is little possibility of a replication under current institutional norms, although two similarly capable researchers have already done "pre-replications" that are extremely relevant as sources of important evidence regarding the hypothesis tested. The balance of the evidence for an effect across the three studies is relatively weak, but only evidence from one study with a significant coefficient is publishable in top journals.

We next address these issues and others relevant to statistical hypothesis tests. There are many other issues that could be addressed, but we concentrate on a vital few since our purpose is to start a conversation rather than to discuss extensively the philosophy or techniques or interpretation of statistics.

DOES A SIGNIFICANT *P*-VALUE HAVE VALUE?

Null Hypothesis Significance Tests (NHSTs) are the core of quantitative empirical research in strategic management and in many other fields of study. The key component of NHSTs is the *p*-value. Particular *p*-values (0.05, 0.01, or 0.001) have been endowed with almost mythical properties far removed from the mundane probabilistic definition.

P-values arising from NHTSs provide no information "regarding the reliability of the research" (Branch, 2014: 257, citing 13 other papers). It is incorrect to interpret *p* as the probability that the null hypothesis H0 is false. Instead, *p* is the probability that the sample value would be at least as large as the value actually observed if the null hypothesis is true (Wonnacott and Wonnacott, 1990: 294), or prob(sample|H0 is true).[1] Sample statistics are about samples rather than populations. However, many mistakenly turn this conditional probability around to say prob(H0 is true|sample). Conditional probabilities cannot be reversed. For example, Branch (2014: 257) notes that prob(dead|hanged) is not the same as prob(hanged|dead).

More generally, consider the truth or falsehood of the following six questions (Gigerenzer, 2004; Haller and Krauss, 2002; Oakes, 1986), assuming you have just found that $p = 0.01$.

1. You have disproved the null hypothesis (that there is no difference in population means).
2. You have found the probability of the null hypothesis being true.
3. You have proved the experimental hypothesis (that there is a difference between the population means).
4. You can deduce the probability of the experimental hypothesis being true.
5. You know, if you reject the null hypothesis, the probability that you are making a wrong decision.
6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99 percent of occasions.

All six statements are incorrect, even though experiments in fields such as psychology have found that students; professors, including those teaching statistics courses; and lecturers routinely believe that at least some of the statements are correct (e.g., Gigerenzer, 2004).

Notice that the true definition of *p*-value permits a conclusion only about the probability of finding a result in a particular sample. This is one reason why "searching for asterisks" (Bettis, 2012) is so problematic. The approach tunes the result to the sample, thereby destroying the whole hypothesis testing logic. Furthermore, the practice is inconsistent with Popper's falsifiability criterion.

DOES SIZE MATTER?

The size of *p*-values is often taken as a measure of the "strength" of the result, where smaller *p*-values are considered stronger evidence. Furthermore, certain *p*-values (0.05, 0.01, and 0.001) are considered "critical" in testing a hypothesis. A value of 0.05 often becomes the breakpoint between truth and irrelevancy. A rigid *p*-value breakpoint is inconsistent with good science. This point leads to a closely related issue—the size of the coefficient tested.

[1] The definition in Wonnacott and Wonnacott (1990) applies to a value in the right-hand tail of the distribution. For a sample value in the left-hand tail, the *p*-value is the probability that the sample value would be at least as small as the value actually observed.

DOI: 10.1002/smj

It is troubling that the strength of a result in terms of economic, behavioral, or practical importance, as indicated by the size of the estimated coefficient, is sometimes ignored if the coefficient is significant. A significant *p*-value without economic, behavioral, or practical importance is in some sense "useless." It seems prudent to consider the material importance of any coefficient as a vital part of the publication process (e.g., McCloskey and Ziliak, 1996; Ziliak and McCloskey, 2004).

REPLICATION IS THE MEASURE OF REPEATABILITY

Replications of statistical studies can take many forms (for a more detailed discussion, see the forthcoming *SMJ* special issue on replication). These include replications that probe the robustness of an original study using a different sample of data in a similar setting (e.g., the same industry) or the generalizability of the results to different settings (e.g., firm size, industry, geography). Publishing only statistically significant results, while not publishing replications or nonresults is inconsistent with the establishment of repeatable cumulative knowledge. One significant coefficient in one study proves little or nothing. Instead, it establishes initial confirming evidence. Similarly, a single replication without statistical significance on the coefficient(s) of interest does not disprove anything. Instead, it establishes or adds disconfirming evidence. Because the nature of statistical testing is probabilistic, we can only make statements about the balance of evidence. It is the balance of evidence that is important to consider.

IS "INTERESTINGNESS" A SCIENTIFIC CRITERION FOR THE VALUE OF RESEARCH?

Throughout the research community of strategic management and much of social science, the term *interesting* has taken on a particular connotation and vital role in establishing the rationale for any particular research study. This largely reflects the massive impact of a single sociology paper, "That's Interesting" by M. S. Davis (1971). It is typical for Ph.D. students in strategic management to encounter this paper and its emphasis on "interestingness" as the primary criterion for research choices many times during a Ph.D. program. The primary word that Davis (1971) uses in describing "interesting" is *counterintuitive*.

If this criterion is correct, then the world has arranged itself so that all phenomena of research importance are counterintuitive. As a theory of the structure of the social universe this seems odd. Indeed, there are many problems that may seem intuitive to at least some scholars, but important to progress in strategic management. For example, it seems intuitive that research and development may affect innovation outcomes—but this relationship clearly merits study.

We suggest that the real issue is to return the word *interesting* to its standard English language meaning of something that you want to learn more about. This includes building cumulative knowledge of strategic management phenomena through replications and publication of nonresults.

WHAT CAN I DO?

As the Co-Editors of *SMJ*, we want to encourage and help strategic management scholars change professional norms related to how we do and interpret statistical research. We all need to work together to solve these problems, and thereby, make the growing knowledge base in strategic management reliable and cumulative. As a first step, there is a compelling need for a broad and deep discussion of these issues among strategic management authors, referees, and editors. We suggest that individual scholars consider taking some of the following actions:

1. Educate yourself and others about the proper use and interpretation of statistics.
2. Schedule a department seminar on this topic.
3. Develop seminar sessions to educate Ph.D. students about the proper use and interpretation of statistics.
4. Organize PDWs and conference sessions about replication, the proper use and interpretation of statistics, or overcoming institutional barriers to repeatable cumulative research.
5. Organize a miniconference on these topics.
6. Engage in replication research. In the near future, *SMJ* will be publishing guidelines for research that replicates important studies.
7. Encourage more top-tier journals to publish replication studies and "nonresults."

DOI: 10.1002/smj

NEW POLICIES AT *STRATEGIC MANAGEMENT JOURNAL*

As a first step, *SMJ* is implementing the following policies, effective immediately:

1. *SMJ* will publish and welcomes submissions of replication studies. Additional guidelines will be provided in the forthcoming *SMJ* special issue on replication.
2. *SMJ* will publish and welcomes submissions of studies with nonresults. These types of studies demonstrate a lack of statistical support in a particular sample for specific hypotheses or research propositions. Such hypotheses or propositions should be straightforward and logical. Studies should be conducted rigorously and assess the robustness of the nonresults, such as robustness to alternative variables, statistical specifications, and estimation methodologies.
3. *SMJ* will no longer accept papers for publication that report or refer to cut-off levels of statistical significance (*p*-values). In statistical studies, authors should report either standard errors or exact *p*-values (without asterisks) or both, and should interpret these values appropriately in the text. Rather than referring to specific cut-off points, the discussion could report confidence intervals, explain the standard errors and/or the probability of observing the results in the particular sample, and assess the implications for the research questions or hypotheses tested.
4. *SMJ* will require in papers accepted for publication that authors explicitly discuss and interpret effect sizes of relevant estimated coefficients.

SUGGESTED READING

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

Branch M. 2014. Malignant side-effects of null-hypothesis testing. *Theory and Psychology* **24**(2): 256–277.

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

Goldfarb B, King A. 2016. Scientific apophenia in strategic management research: significance tests & mistaken inference. *Strategic Management Journal* **37**(1): 167–176.

Kerr NL. 1998. HARKing: Hypothesizing after results are known. *Personality and Social Psychology Review* **2**(3): 196–217.

Matulsky H. 2014. *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. Oxford University Press: New York, NY. (Part D, titled, “P Values and Statistical Significance” is superb, very readable and immediately applicable. Also, see Chapter 45, “Statistical Traps to Avoid.”)

Ziliak SL, McCloskey DN. 2004. Size matters: the standard error of regressions in the American Economic Review. *Journal of Socio-Economics* **33**: 527–546.

REFERENCES

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

Branch M. 2014. Malignant side-effects of null-hypothesis testing. *Theory and Psychology* **24**(2): 256–277.

Davis M. 1971. That’s interesting! towards a phenomenology of sociology and a sociology of phenomenology. *Philosophy of the Social Sciences* **1**: 309–344.

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

Goldfarb B, King A. 2016. Scientific apophenia in strategic management research: significance tests & mistaken inference. *Strategic Management Journal* **37**(1): 167–176.

Haller H, Krauss S. 2002. Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research Online* **7**(1): 1–20. Available at: http://www.mpr-online.de (accessed 6 November 2015).

Horton R. 2015. Offline: what is medicine’s 5 sigma. **385**. Available at: http://www.thelancet.com.

Ioannidis PAJ. 2005. Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* **294**(4): 218–228.

McCloskey DN, Ziliak ST. 1996. The standard error of regression. *Journal of Economic Literature* **24**: 97–114.

Oakes M. 1986. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. Wiley: New York, NY.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* **349**: 943.

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

Wonnacott TH, Wonnacott RJ. 1990. *Introductory Statistics for Business and Economics*. Wiley & Sons: New York, NY.

Ziliak SL, McCloskey DN. 2004. Size matters: the standard error of regressions in the American Economic Review. *Journal of Socio-Economics* **33**: 527–546.