Journal of Management Special Issue on Robust and Reliable Research Practice
Lead-in Commentary

EMBRACING ROBUSTNESS AND RELIABILITY IN THE SCIENCE OF ORGANIZATIONS

Gwendolyn K. Lee
Gwendolyn.lee@warrington.ufl.edu

Mo Wang
Mo.wang@warrington.ufl.edu

Warrington College of Business
University of Florida
1384 Union Road, Bryan Hall 100
Gainesville, FL 32611

Acknowledgments: We thank David Allen, Editor-in-Chief, Journal of Management, for commissioning the Special Issue. We also thank the National Science Foundation for funding the Workshop on Promoting Robust and Reliable Research Practice in the Science of Organizations held at the University of Florida (Award ID: SES1743044).

EMBRACING ROBUSTNESS AND RELIABILITY IN THE SCIENCE OF ORGANIZATIONS

ABSTRACT

The science of organizations increases its credibility when it embraces research with an explicit focus on robustness and reliability. This special issue of curated commentaries recommends and illustrates how to incorporate robust and reliable research practice in organizational research. Together, these commentaries help researchers make contributions to improving every step of the research trajectory—theory development, methodology, and the process of quality control through peer review.

Keywords: Robustness, Reliability, Reproducibility, Replicability, Generalizability, Science of organizations

The science of organizations increases its credibility when it embraces research with an explicit focus on robustness and reliability. Questionable research practices that operate in the ambiguous space between what one might consider best practices and academic misconduct, however, appear to occur at a non-ideal frequency (Banks *et al*., 2016). Increasing concerns about credibility crisis among management scholars (e.g., Aguinis *et al*., 2017; Bergh *et al*., 2017a; Karabag & Berggren, 2012) motivate this special issue to curate commentaries that recommend and illustrate how the studies in management journals can become more credible.

While many thought leaders in the field of management and further afield have provided broad, practical and evidence-based recommendations for research practices of good science (e.g., Banks *et al*., 2016; Bergh *et al*., 2017a; Bettis *et al*., 2016; Munafo *et al*., 2017; Nelson *et al*., 2018; Wright, 2016), this special issue offers a complementary approach. Many extant recommendations seek to increase the credibility of academic research by targeting the way scientists collect and analyze data and focusing on methodological improvements. By expanding the focus beyond methodology, the set of commentaries in this special issue engages in a deeper conversation about robustness and reliability in the current research trajectory that is often adopted by researchers. The research trajectory typically starts with theory, followed by implementation and then empirically established solutions. Robust and reliable research practice augments the research trajectory by adding more clarity on the structures, mechanisms, and boundary conditions postulated in a theory, higher precision for theory testing, and an accumulation of solid empirical foundation. As a collection, the set addresses how to incorporate robust and reliable research practice in every step of scientific research—theory development, methodology, and the process of quality control through peer review.

Shaver (in this issue) explains how to design research that improves the ability to isolate the underlying cause of a relationship between two variables – should one exist. Identifying the causal mechanism that creates a relationship between two variables is an important yet daunting goal, particularly in the study of strategy and organizations. Rather than blindly importing solutions from related social science disciplines, Shaver discusses how to define our own path forward. The solution he proposes is to advance causal identification of important questions through a cumulative body of research. Generating a cumulative body of research that systematically advances causal identification means many related studies are required to provide insight into the causal mechanisms. These studies make contributions by ruling out competing causal mechanisms that previous research has been unable to rule-out, or by isolating alternative causal mechanisms that have not been previously considered. The studies also contribute by accurately and transparently describing what was done and what is needed to help future research establish causality.

Shaver's proposal has several implications for improving every step of the research trajectory. For theory development, one implication is that establishing causal mechanisms would be the focal pursuit, rather than advancing a new theory or proposition. For methodology, one implication is that developing a research design and empirical strategy for causal identification would play a key role, which is often missing in many empirical papers that move from hypotheses straight to descriptions of the data and estimation techniques used to test the hypotheses. For the process of quality control through peer review, one implication is that devising novel research designs with rigorous tests of theories would constitute an important contribution. Yet, improving the ability to identify causality does not mean abandoning the

important but messy questions central to the study of strategy and organizations – the key is to take a systematic approach to achieve causal identification for messy questions.

Csaszar (in this issue) shows how formal models—a research method that uses mathematical relationships to provide an abstraction of a phenomenon under study in a way that allows for thinking about the phenomenon using symbolic reasoning and computation—can improve theory development and theory testing in the science of organizations. Theory development is improved when the implications of a set of assumptions are logically derived with increased theoretical precision. Such derivation is particularly useful when the phenomenon under study entails interaction and aggregation effects, boundary conditions, dynamics, and multiple levels of analysis. It is also useful when no ideal data is available. In addition, the increased theoretical precision in formal models simplifies the process of designing empirical tests, whereas verbal theories are often less clear about how to test a theory. Formal models also improve theory testing by stimulating novel ways of constructing a measure and generating testable predictions.

Csaszar's guide on formal models suggests an alternative and faster cycle of theory development and theory testing in the research trajectory. Formal models can accelerate the speed at which organization theory advances, by providing clear definitions, measurable constructs, and transparent mechanisms. The process of quality control through peer review is enhanced, because the interpretation of the formal model does not depend on a subjective system. The interpretation would be the same regardless of who does the interpreting.

Xu, Zhang, and Zhou (in this issue) address the robustness of findings from studies that collect, process, and analyze naturally occurring digital footprints of human activities. Rather than following an explicit research design, those studies use data that have been recorded

continuously by ubiquitous sensors, mobile applications and online social networks, or even generated fictitiously by automated software (i.e., "bots") that pretends to be human users. Xu *et al*. point out several issues about data generation that threaten the validity of inferences drawn from such studies, particularly the inferences made about constructs (i.e., construct validity) and causal relationships (i.e., internal validity).

One of the issues is researchers' motivation to cherry-pick a parameter/procedure space for an information-extraction algorithm so as to generate measures of a construct that will result in a false positive by chance. While pre-registration has been suggested as a remedy for *p*-hacking,[1] requiring researchers to pre-register the algorithmic parameter/procedure before collecting the data poses technical challenges that are unique to this type of data. The machine learning research community has not yet established common guidelines on how to judge the appropriateness of the parameter/procedure-tuning process (i.e., how much tuning is deemed too much). Another issue is the black-box design of digital platforms that can introduce unobserved links between variables such that the inferred causal relationships between the variables are simply the platform's hidden design. The hidden design, such as the recommendations that are generated based on a user's past browsing history, is often proprietary and therefore not transparent to researchers. The platform may make unannounced changes that are difficult to detect.

Bliese and Wang (in this issue) make clear that statistically significant results have different long-term probabilities of being significant, and address the uncertainty associated with whether an independent study using similar measures and a similar design would find similar effects. Commonly reported statistics such as *z*, *t* and *p* values from a study convey the long-term

[1] P-hacking refers to modifying statistical models over the same data until statistically significant results are found (Nelson *et al*., 2018).

probability of finding statistically significant effects, but a transformation of those commonly reported statistics is required to move beyond dichotomous tests of statistical significance. The commentary illustrates that such a transformation, which can be achieved via using the non-parametric bootstrap or formula-based approaches, offers an estimate of the long-term probability of significance (i.e., the "observed power" or "post-hoc power"; Hoenig & Heisey, 2001; Yuan & Maxwell, 2005), given the characteristics (and limitations) of a specific sample and the statistical model being used. The observed power provides a way for researchers to report the level of uncertainty associated with the finding and the potential variability in results. Reporting the variability in results presents a sharp contrast from following a practice commonly adopted in the conventional Null Hypothesis Significance Testing (NHST) paradigm of point estimate (see Schwab *et al*., 2011; Gelman, 2018 for problems with the NHST), which is using a threshold to dichotomize a result as statistically significant versus not statistically significant.

Bliese and Wang recommend authors to report observed power to help counter origination bias (i.e., how much a single-study finding should be viewed as solid or sacred) and other biases tied to misunderstanding the variability inherently associated with "statistically significant" findings. Indeed, effects and patterns can and do change over time, and they can vary across samples. "If effects are different in different places and at different times, then episodes of nonreplication are inevitable, even for very well-founded results" (Gelman, 2015: 633). One consequence of origination bias is that researchers have little incentive to replicate novel findings. Bliese and Wang argue that by estimating the long-term probability of significance, authors can establish circumstances where replicating a finding is less likely to make a substantial contribution.

Bliese and Wang also recommend the reviewers and editors to use the information conveyed by observed power to help authors avoid over-selling their findings by assuming those findings' robustness. Further, using this information, it will also be clear that the strict adherence to dichotomous interpretations of *p* values makes little sense, because the long-term probabilities of a finding with a *p* value of .044 (significant) is basically identical to the long-term probabilities from a finding with a *p* value of 0.055 (non-significant).

In summary, the collection of commentaries in this special issue provides a guide, helping researchers improve their abilities to increase credibility in every step of organizational research. Beyond what are covered in the collection, we also would like to emphasize the importance for increasing the transparency concerning the measures, manipulations, and exclusions, as well as how the authors determined their sample sizes and stopping rules, in research reporting in the science of organizations. When multiple data processing and data analytic decisions were explored, how the decisions were made and narrowed down needs to be fully disclosed in the manuscript. The methods, procedures, and computational steps including programing code that were taken to arrive at analyzed datasets and results also need to be fully disclosed. If subsequent researchers gain access to the initial study's data, they should be able to produce the same results. In addition, descriptive and correlational statistics need to be disclosed sufficiently such that reviewers and subsequent researchers can recreate a data set that is statistically equivalent to the initial study's and use them to retest the study's reported models. Subsequent analyses will be identical whether using the data matrix or the complete raw data file itself (see Shaver, 2005; Boyd et al., 2010; Bergh et al., 2017b for illustrations within the management literature). Using the disclosed statistics, a set of systematic error-detecting checks such as the "Red Flag" tests recommended and demonstrated by Bergh et al. (2017b) can be used

by editors to verify that the statistical results reported in a manuscript are accurate, when the manuscript reaches the conditional acceptance point of evaluation. A standard statement, such as the one endorsed by the Center for Open Science, can be used by reviewers to make requests for disclosure of data collection, analysis, and statistics.

Finally, it is important to note that the recommended research practices that we curate in the special issue encompass a shift from null hypothesis significance testing of point estimate toward understanding the variability in effect sizes across studies. Shifting toward assessing the variability in effect size, by contrast, focuses subsequent studies on uncovering a continuous distribution on the magnitude of effects rather than assessing the probability of the sharp point null hypothesis of zero effect and zero systematic error (Gelman & Carlin, 2017).[2] The shift signifies that we accept with humility that we gain knowledge without the certainty we might like, as we learn to embrace the uncertainty in discovering and building repeatable, cumulative, and causal research knowledge.

[2] Shifting toward assessing the heterogeneity in effect size is a general recommendation, not specific to Bayesian inference. “In small-sample studies of small effects, often all that a good Bayesian analysis will do is reveal the inability to learn much from the data at hand. In addition, little is gained from switching to Bayes if you remain within a traditional hypothesis-testing framework. We must move beyond the idea that effects are “there” or not and the idea that the goal of a study is to reject a null hypothesis” (Gelman, 2015: 640).

REFERENCES

Aguinis, H., Cascio, W. F. & Ramani, R. S., 2017. Science's reproducibility and replicability crisis: International business is not immune. *Journal of International Business Studies*, 48(6): 653-663.

Banks, G. C., O'Boyle Jr., E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., & Adkins, C. L. 2016. Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, 42(1): 5-20.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. 2017a. Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization*, *15*(3): 423-436.

Bergh, D. D., Sharp, B. M., & Li, M. 2017b. Tests for identifying "Red Flags" in empirical findings: Demonstration and recommendations for authors, reviewers, and editors. *Academy of Management Learning & Education*, 16(1): 110-124.

Bettis, R. A, Helfat, C. E., & Shaver, J. M. 2016. The necessity, logic, and forms of replication. *Strategic Management Journal,* Special Issue: Replication in Strategic Management, 37(11): 2193–2203.

Bliese, P. D., & Wang, M. This issue. Your Study Results Provide Information about the Long-Term Probability of Finding Significant Effects: Let's Report this Information. *Journal of Management*,.

Boyd, B. K., Bergh, D. D., & Ketchen, D. J., Jr. 2010. Reconsidering the reputation-performance relationship: A resource-based view. *Journal of Management*, 36(3): 588–609.

Csaszar, F. This issue. *Certum Quod Factum*: How Formal Models Contribute to the Theoretical and Empirical Robustness of Organization Theory. *Journal of Management*,.

Gelman, A. 2015. The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*, *41*(2): 632-643.

_____. 2018. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, 44(1): 16-23.

Gelman, A., & Carlin, J. 2017. Some natural solutions to the p-value communication problem—and why they won't work. *Journal of the American Statistical Association*, 112(519): 899-901.

Hoenig, J. M., & Heisey, D. M. 2001. The abuse of power. *The American Statistician,* 55: 19-24.

Ioannidis, J. P. A., & Trikalinos, T. A. 2007. An exploratory test for an excess of significant findings. *Clinical Trials*, 4(3): 245–53.

Karabag, S. F., & Berggren, C. 2012. Retraction, dishonesty and plagiarism: Analysis of a crucial issue for academic publishing and the inadequate responses from leading journals in economics and management disciplines. *Journal of Applied Economics and Business Research*, 2(3): 172-183.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. 2017. A manifesto for reproducible science. *Nature Human Behaviour*, 1(0021): 1-9.

Nelson, L. D., Simmons, J., & Simonsohn, U. 2018. Psychology's renaissance. *Annual Review of Psychology*, 69: 511-534.

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. 2011. Perspective—researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science*, *22*(4): 1105-1120.

Shaver, J. M. This issue. Causal Identification Through A Cumulative Body of Research in the Study of Strategy and Organizations. *Journal of Management*,.

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

Wright, P. M. 2016. Ensuring research integrity: An editor's perspective. *Journal of Management,* 42(5): 1037-1043.

Xu, H., Zhang, N., & Zhou, L. This issue. Towards Robust Research Using Organic Data. *Journal of Management*,.

Yuan, K., & Maxwell, S. 2005. On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics,* 30: 141-167.