SURPRISES, CONFLICTING FINDINGS, OR QUESTIONABLE RESEARCH PRACTICES? A METHODOLOGY FOR EVALUATING CUMULATIVE EMPIRICAL ANALYSES AND REPLICATION STUDIES

Gwendolyn K. Lee

Chester C. Holloway Professor & Recipient of the University Term Professorship
Warrington College of Business
University of Florida
1384 Union Road, Bryan Hall 100
Gainesville, FL 32611, USA.

ORCID: https://orcid.org/0000-0003-4276-7982
E-mail: glee@alum.mit.edu

Acknowledgments: I thank Jeff Reuer, Co-Founding-Editor of *Strategic Management Review*, for championing the integration of our research efforts and the construction of a robust, cumulative body of knowledge as key opportunities facing the field. I, as the principal investigator, also thank the National Science Foundation for funding the Workshop on Promoting Robust and Reliable Research Practice held at the University of Florida (Award #SES1743044 Division of Social and Economic Sciences, Science of Organizations Program). A website archiving the content generated from the workshop is hosted at https://warrington.ufl.edu/reliable-research-in-business/.

SURPRISES, CONFLICTING FINDINGS, OR QUESTIONABLE RESEARCH PRACTICES? A METHODOLOGY FOR EVALUATING CUMULATIVE EMPIRICAL ANALYSES AND REPLICATION STUDIES

Abstract

Critiques about the research practices that the scholars in strategic management engage in have called out that the field of strategic management appears vulnerable to a credibility crisis. As the field accumulates discrepancies between an initial observation and subsequent observations about a theoretical expectation, how do we know that the discrepancies are surprises, conflicting findings or questionable research practices? Questionable research practices that operate in the ambiguous space between what one might consider best practices and academic misconduct alert the research community to confront the discrepancies. Yet, the field does not have a methodology for diagnosing the root causes of discrepancies in cumulative empirical analyses. In the current article, we propose a methodology that uses abductive reasoning in the evaluation of discrepancies. Abductive reasoning is a process for reacting to discrepancies through model reformulation, revision of hypotheses, and addition of new information. The proposed methodology may aid not only authors, but also journal editors and reviewers, in evaluating discrepancies and assessing the merits of replication studies.

Keywords: Science of science; Questionable research practices; Robustness; Replication studies; Recommendations for peer review and editorial guidelines.

Introduction

Critiques on the replicability of scientific research have alerted possible errors and potentially false knowledge in many fields (e.g., the replicability of experimental studies in the social sciences by Camerer et al., 2016; Camerer et al., 2018; Open Science Collaboration, 2015). The alerts raised in the critiques have generated heated debates and contentious disagreements in some fields.[1] In the field of strategic management, based on the critiques of Aguinis and Solarino (2019), Bergh et al. (2017a), Goldfarb and King (2016) that assess whether empirical findings that were reported in a top journal's articles can be reproduced based on the articles' published methods and data, the field appears vulnerable to a credibility crisis.

While a credibility crisis looms in our field, there is an increasing acceptance of replication studies (Ethiraj, Gambardella & Helfat, 2016). Studies that are designed to replicate previously published research provide critiques of the discrepancies in cumulative empirical analyses—the discrepancies between an initial observation and subsequent observations about a theoretical expectation (e.g., Goldfarb & Yan, 2021, on a theoretical expectation that, when organizations are recognized as legitimate players in a category, they perform better). As the field accumulates discrepancies, how do we as a community of scholars evaluate discrepancies? Whereas a meta-analysis serves as a science of science approach for integrating a collection of findings, an evaluation of discrepancies in cumulative empirical analyses and replication studies is a critique that

[1] As an example, Gilbert et al. (2016) criticized three statistical errors in Open Science Collaboration (2015), and argued that, when the OSC results are corrected for error, power, and bias, the data would lead to a conclusion that is opposite from the OSC's finding of a surprisingly low reproducibility in psychological science. Anderson et al. (2016: 1037-c) responded to Gilbert et al.'s criticism and suggested that Gilbert et al.'s "very optimistic assessment is limited by statistical misconceptions and by causal inferences from selectively interpreted, correlational data." Anderson et al. maintained that both optimistic and pessimistic conclusions about reproducibility are possible, and neither are yet warranted.

alerts possible errors and potentially false knowledge. Critiques of discrepancies, as explained by Gelman (2018), exert continuous pressures on authors and editors against publishing, in the first place, unfounded claims and serious flaws. While critiques serve as a mechanism of quality control for robust and reliable research, we need evaluation criteria to tell apart three types of discrepancies: (1) surprises; (2) conflicting findings; and (3) questionable research practices.

In the current article, we classify into three types the discrepancies between an initial observation and subsequent observations about a theoretical expectation. For each type, we provide an example of a critique that has been offered about such discrepancies. We then offer abduction as a form of reasoning in the evaluation of discrepancies in cumulative empirical analyses and replication studies. Abductive reasoning is a process for reacting to discrepancies through model reformulation, revision of hypotheses, and addition of new information. Abductive reasoning consists of three modes of evaluation—descriptive, prescriptive, and normative—according to a typology of criteria that Mantere and Ketokivi (2013) developed for the evaluation of scientific reasoning in organization research. We submit that these three modes form the basic evaluation criteria for classifying and telling apart the discrepancies.

Abductive reasoning in the evaluation of discrepancies has implications for strategic management. The evaluation can usefully return our attention to the canonical research questions of strategic management. As an example, a replication study on the academic research on diversification discount by Chang, Kogut and Yang (2016: 2255) observed that, "[d]iversification is one of the perennial themes in the history of strategy research […] It is surprising that studies find no value, indeed value destruction, to

industry or global diversification." Their replication study showed results that contradicted past research in which global diversification had been reported to destroy firm economic value. Chang et al.'s critique, as will be explained later in the current article, uses abductive reasoning in their evaluation of the discrepancies in what has been reported in the academic literature. Importantly, it contributes to the managerial practice of multinational corporations by constructing a credible and useful body of knowledge around one of the canonical questions: Is there value in being diversified?

Surprises, Conflicting Findings, or Questionable Research Practices?

We showcase three discrepancies in the literature on corporate governance where replication studies and the critiques about such discrepancies have been published. The first one is a discrepancy in the findings about an association between governance mode and performance outcome, where a theoretical expectation based on the traditional principal-agent model on incentives is that franchising and vertical integration would differ in performance outcomes. Franchising is a governance mode where a franchisor creates a product, business plan, and trademarks and sells the right to open a branded store to a franchisee. Whereas the empirical evidence in the extant literature has been consistent in suggesting that differences in governance mode lead to differences in performance, Kosová, Lafontaine and Perrigot (2013) find that franchising has no effect on performance in their data on the operations of a large multi-chain hotel company. The critique that the authors offer regarding the discrepancy is that the differences in performance outcomes between franchised and corporate hotels become statistically insignificant once the choice of governance mode is endogenized. In the raw data, they

find significant differences—higher prices and lower occupancy rates among franchised than among corporate hotels. However, when the governance mode is not constrained and can be chosen to optimize performance (larger revenues per room, higher occupancy rates, or better prices), they argue it is not a surprise to find that, after controlling for factors that affect the choice of governance mode, franchising itself does not give rise to different performance outcomes (p.1319).

The second one is a discrepancy in the findings about an association between corporate social and financial performance, where six studies using the same data sources and similar methods deliver conflicting findings and interpretations. These studies report divergent estimates suggesting an association that may be positive and linear, spurious, limited to particular scales, moderated, U-shaped, or conditional on certain financial measures (see Appendix 1 in Berchicci and King, 2021, for more details). A critique offered by Berchicci and King (2021) regarding this set of conflicting findings is that 'many quasi-replications use the same data sources and high-level empirical designs, yet still differ in many ways, both seen and unseen by readers. They may measure variables differently, or assume different functional forms, or define samples using different processes. Readers who wish to interpret conflicting results from such studies must try to discern the cause of conflicting "findings." How would estimates have changed if different empirical assumptions had been used?' Their critique further points to a source of model uncertainty that juxtaposes between- versus within-firm variations: firms with high social performance tend to have higher financial performance, but improvement in social performance seems, if anything, to reduce financial performance.

The third one is a discrepancy in the findings about an association between CEO gender and executive compensation, where Hill, Upadhyay and Beekun (2015) reported that female CEOs receive greater compensation than male CEOs, a finding that runs counter to common wisdom that the gender pay gap in the labor market favors men over women. In the critique offered by Gupta, Mortal and Guo (2018) regarding the discrepancy, they state that their direct replication fails to find reliable evidence that could support Hill et al.'s finding. What explains the difference in findings between Hill et al. and Gupta et al.? Gupta et al. (2018: 2045) speculated that their inability to reproduce Hill et al.'s sample using the same data sources or to replicate Hill et al.'s finding using multiple samples (as explained in Gupta et al.'s Table 2) could be due to some unstated aspect(s) of Hill et al.'s data collection and analyses not readily obvious from the methodological description.[2] It is important to note that the journal editor who accepted Gupta et al.'s (2018) article published in an online appendix the computer codes for interested readers to use with SAS and STATA software (footnote 3, pp.2041).

Suppose the single objective of peer review and editorial leadership is to develop cumulative knowledge that is credible. How would authors and their best critics evaluate a finding where a statistically and economically significant correlation diminished and disappeared as the model specification changed (such as adding control and instrumental variables, modifying functional form, using different econometric methods)? How would authors and their best critics evaluate a finding where a statistically and economically significant correlation didn't exist but appeared as the model specification changed? How

[2] Gupta et al. (2018: 2045) declared that, "Given the observation that replication of significant results would likely fail in about half of the published strategy studies (Goldfarb & King, 2016), our inability to replicate HUB's finding about gender differences in CEO compensation should not be very surprising (though it is concerning)."

would the evaluation differ if the authors revealed transparently the garden of forking paths (Gelman & Loken, 2014 referring to the space of choices and assumptions that are made by empirical researchers in their search of statistical significance) and reported a visualization of the garden with an epistemic map showing the uncertainties in the findings (King, Goldfarb & Simcoe, 2021; Simonsohn, Simmons & Nelson, 2020)? In addition to mapping the garden and reporting epistemic uncertainties, what other merits are necessary and sufficient for authors, reviewers, and editors when diagnosing the root causes of discrepancies in cumulative empirical analyses? Our field has not yet established a standardized methodology to answer these questions.

Abductive Reasoning in the Science of Science

As a first step toward answering these questions, we offer abduction as a form of reasoning for evaluating discrepancies in cumulative empirical analyses and replication studies. Abductive reasoning as a form of scientific reasoning was first analyzed by Charles Sanders Peirce, who described a process for reacting to surprises in the following way: "The surprising fact, C, is observed. But if A were true, C would be a matter of course. Hence, there is reason to suspect that A is true" (Peirce, 1934: 117). Suspicion about A opens the door and encourages further discovery and investigation through model reformulation, revision of hypotheses, and addition of new information. This process of discovery and investigation—generating and revising models, hypotheses, and data analyzed—is abduction. Peirce characterized abduction, which is distinct from induction or deduction, as a form of inference that moves descriptions of the world forward rather than just confirming or falsifying hypotheses (cf. Popper, 1959, 1963).

The "official Peirce abduction schema" is often referred to as retroduction, a form of reasoning that extracts from data an explanation that accounts for particular observations (Schurz, 2008: 206). Retroduction was advocated by Hanson (1961: 85-88), who followed Peirce (1878a) and influenced Simon (1968: 339, 443, & 456). Retroduction derives generalizations from data by observing raw data and concluding that the observed data can be described adequately by a theoretical model that explains causes and causal mechanisms. Simon (1968: 443) on judging the plausibility of theories offered the rank-size distribution of city populations in the United States as an example when explaining retroduction as a process of inference from the facts. A simple generalization to some degree of approximation is that size varies inversely with rank. If the generalization fits the facts, then a log-log plot of the data would suggest points falling on a straight line with a slope of minus one. He argued that "the standard statistical tests of hypotheses are inappropriate" because "the theory of statistical tests gives us no real help in choosing between an approximate generalization and an invalid one" (pp.440, 443). "It is quite easy to find data that are quite curvilinear to the naked eye (see fig. 3)" which is the rank-size distribution of cities in Austro-Hungarian Empire, 1910 and in Austria, 1934 (pp.444, 446). "We may therefore find the evidence unconvincing that the phenomena are 'really' linear in the limiting cases. The phenomena are not striking enough in this respect to rule out coincidence and chance. Should we believe the data to be patterned?" (pp. 444).

As shown in the example that Simon illustrated, retroduction generates a plausible explanation with a mechanism that specifies the conditions under which linearity should hold most exactly and the slope should most closely approximate to minus one. When

these conditions do not hold, as in the examples that Simon gave such as Austria after World War I, it is not a surprise that the data are not patterned as a straight line with a slope of minus one. Through abductive reasoning, scientists generate new hypotheses such as the mechanism and boundary conditions that Simon conjectured. The new hypotheses are meant to account "for surprises and unmet expectations" (Locke, 2010) by moving from local observations in a particular situation to untested explanations.

By contrast, deductive reasoning and inductive reasoning are concerned with general applicability. Through deductive reasoning, one draws a conclusion about the particular based on the general, whereas, through inductive reasoning, one moves from the particular to the general. Abductive reasoning, as reckoned by Heckman and Singer (2017), generates defensible explanations for surprising phenomena. It looks for consilience across bodies of evidence and across studies instead of reporting results in isolation. It diminishes the value of any particular study but encourages further exploration and testing on multiple sources of evidence. It produces more true knowledge and fewer statistical artifacts arising from particular sequences of choices of analyzing datasets. It is a strategy for growing knowledge and not for pretending to have it (ibid: 301). However, there is no established practice or formal guidelines for taking the next step and learning from empirical surprises (Heckman & Singer, 2017: 298).

Abductive reasoning is particularly good for exploring and discovering explanations in research contexts where the empirical surprises might come from (Behfar & Okhuysen, 2018: 329). "Reliance on abductive reasoning as a primary focus of a manuscript is reasonable here because of the need to explore and discover a new and plausible explanation, one that restores theoretical coherence in light of empirical reality.

Here, the researcher can drop the tools (and rules) of hypothetico-deductive (H-D) inquiry and create a manuscript that relies on abductive inquiry to put forth a new set of plausible explanations." As suggested by Benfar and Okhuysen (2018), Heckman and Singer (2017) and Simon (1968), abductive reasoning can be used to generate new and plausible explanations.

Adding to these suggestions, we submit that abductive reasoning can be used in the evaluation of cumulative empirical analyses and replication studies when a body of knowledge exhibits empirical inconsistencies, contradictions, or discrepancies. One exemplar in the use of abductive reasoning in a critique of discrepancies is Chang, Kogut and Yang (2016). The discrepancies start with the observation that two articles on diversification discount use similar data and similar variable specifications following Berger and Ofek (1995), but differ in their econometric specification. Denis, Denis and Yost (2002) found a global diversification discount without correcting for selection, but Campa and Kedia (2002) found that the industry diversification discount disappeared or reversed once the econometric specification accounted for selection. Chang et al. (2016) conducted a replication study to examine what happens to the global diversification discount once accounting for self-selection. They found that on controlling for self-selection, the incremental value of global diversification turned from a discount to a premium for the time period 2005–2011. Abductive reasoning is used in not only the diagnosis of self-selection bias as a cause of why their finding contradicts past research, but also the generation of a theoretical mechanism to explain why they also found that, whereas the premium emerged during the period of the 2008–2009 financial crisis, the

period 2005–2007 showed no premium. The mechanism is a new and plausible explanation that accounts for the discrepancy between the two time periods.

Three Modes of Evaluation

For evaluating cumulative empirical analyses and replication studies, we propose three modes of evaluation and their implementations in research practice. The abductive reasoning in a critique of discrepancies consists of three modes of evaluation—normative, descriptive, and prescriptive. The normative mode is based on the selection of the "best explanation" from a set of competing explanations (Ketokivi & Mantere, 2010: 330). The role of normative evaluation is to ensure epistemological rigor, best described as resilience and restraint stemming from appreciating the unavoidable incompleteness of our knowledge claims (Mantere & Ketokivi, 2013: 74). By contrast, the descriptive mode is founded on the transparency of the explanations considered. The role of descriptive evaluation is to provide transparency—to reveal the local aspects of reasoning—and call for the disclosure of cognition in all its idiosyncrasy (ibid: 75). In comparison, the prescriptive mode places an expectation of compliance to local epistemic values in selecting one explanation over the others. The role of prescriptive evaluation is for a scholarly community to assess the credibility of knowledge claims according to the methodological considerations and preferences of the community (ibid: 74).

The Normative Mode of Evaluation

In selecting the "best explanation" such as the root cause from a set of competing explanations to account for the discrepancies between an initial observation and subsequent observations about a theoretical expectation, a normative ideal strives for a

tight connection among theory, measurement, and data. The "best explanation" is selected by the researchers based on pragmatic virtues, such as usefulness, simplicity, conservativeness, or interestingness, which is the potential of an explanation to answer open questions or create new ones (Ketokivi & Mantere, 2010: 330, 219).

We submit that the abductive reasoning in a critique of discrepancies is normative when the evaluation identifies mechanisms and boundary conditions that may lead to deeper and richer theories, while acknowledging that each observation, initial or subsequent, is specific to its context/population examined. The discrepancies may reflect the fact that the true effect size varies across studies. A moderator, for instance, could be a previously unknown factor (e.g., populations, time periods, organizations, geographical areas, measurement instruments, etc.) that can be hypothesized in the critique to account for the discrepancies (see Van Bavel et al., 2016 for an example of exploring contextual sensitivity in scientific reproducibility). The discrepancies may also result from the variation across studies and papers (including differing degrees of variation across various dependent measures) that incorporate new data and other departures from the initial observation such as alternative measures, models, and methods.

Discrepancies are not unique to a particular type of methodological approach, whether the empirical findings were obtained with quantitative, qualitative, or mixed-methods approaches. The use of abductive reasoning in accounting for the discrepancies, however, is different when the discrepancies are among studies that use qualitative and mixed-methods approaches, compared to when the discrepancies are among studies that use quantitative approaches. When the discrepancies are among studies that use qualitative and mixed-methods approaches, one explanation to account for them is the

idiosyncratic context. In the normative mode, a critique can use the richness of the context (e.g., unique organizational features and environmental characteristics) to make sense of the surprising and inconsistent observations. Indeed, the focus on context has long been recognized to offer great explanatory power and generate novel insights and mechanisms in qualitative research, because qualitative research excels at realism (McGrath, 1981).

We further submit that the abductive reasoning in a critique is normative when the evaluation adopts a multilevel analysis (conceptually and/or statistically) that explicitly recognizes that the discrepancies may emerge from the covariation induced by the fact that observations are nested within, for example, papers, studies, groups of subjects, and study conditions (see McShane & Böckenholt, 2018 for an example of a multilevel analysis in the science of science). The nesting of a multilevel model can handle complex interactions by incorporating multiple levels of uncertainty that arise from contexts or populations different from the initial observation (Gelman & Hill, 2007; Gelman et al., 2013; Gelman, 2015).

Moreover, we submit that the abductive reasoning in a critique is normative when the evaluation examines the heterogeneity in effect size. Statistically significant but unreplicable results can be seen as arising from varying treatment effects and situation-dependent phenomena (Gelman, 2015: 640). Findings can vary across studies due to research design artifacts—not only sampling error but also inconsistent construct validity, differences in measurement error, range restriction, dichotomization of continuous variables, and coding and transcription errors across studies (Schmidt & Hunter, 2014). At least two sources of error—sampling error (in cases where subsequent studies use new

samples) and measurement error (in cases where subsequent studies use a measure with different reliability)—can contribute to the discrepancies between an initial estimate and subsequent estimates. Using a simulation, Stanley and Spence (2014) show that measurement error alone generates a high variability in the discrepancies, and when both measurement error and sampling error are present, the variability in the discrepancies increase further. They also report that, as sample size and the reliability in measurement increase, the variability in the discrepancies decreases. These alternative explanations as root causes of the discrepancies—moderators, multilevels, and research design artifacts—are untested hypotheses, which further observations can augment.

The Descriptive Mode of Evaluation

A descriptive evaluation calls for the disclosure of cognition in all its idiosyncrasy, and therefore, a critique's abductive reasoning is descriptive when the evaluation provides the transparency of the selection between alternative explanations. An important disclosure in the science of science is the transparency in researcher degrees of freedom. Researcher degrees of freedom refer to undisclosed flexibility in exploring various analytic alternatives as researchers search for a combination of data collection and analysis that yields 'statistical significance' (Simmons et al., 2011: 1359; cf. King, Goldfarb, & Simcoe, 2021).

Transparency in the reporting of researcher degrees of freedom is necessary for a critique to analyze what was done versus what was disclosed. Incomplete reporting practices, disclosure errors, and possible opportunism limit the reproducibility of most studies (Bergh et al., 2017a). Even for meta-analyses in the science of science, despite

reporting guidelines such as the Quality of Reporting of Meta-Analyses statement (Moher et al., 1999), the checklist for the Meta-Analysis of Observational Studies in Epidemiology (Stroup et al., 2000), and the Meta-Analysis Reporting Standards (MARS; APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008), Aytug et al. (2012) found a considerable amount of variability and inadequacy in the transparency and completeness of reporting in regards to the methodological choices and judgment calls exercised by meta-analysts. The variability and inadequacy add to other sources of distortion to the results of meta-analyses, including questionable research practices used in the studies (Bakker et al., 2012).

We submit that the abductive reasoning in a critique is descriptive when the evaluation compares the measures, manipulations, and exclusions in the empirical research that the critique evaluates. In generating an explanation to account for the discrepancies, the descriptive mode can hypothesize how the decisions about the researcher degrees of freedom were made and narrowed down, including the methods, procedures, and computational steps such as programing code. They can also hypothesize how the authors of the publications being scrutinized may have determined their sample sizes and stopping rules. In addition, they can hypothesize what steps the authors may have taken in arriving at the analyzed datasets and the reported results.

Moreover, we submit that the abductive reasoning in a critique is descriptive when the evaluation explores alternative data analytic specifications. The transparency in alternative data analytic specifications reveals how much the conclusions change because of arbitrary choices in data construction and shows which choices are most consequential

in the robustness of the findings (Steegen et al., 2016). The descriptive mode of evaluation can incorporate a multiverse analysis, which is closely related to the idea of a garden of forking paths in data analysis (Gelman & Loken, 2014; see also Leamer, 1978). A multiverse analysis, as demonstrated by Steegen et al., generates alternative data analytic specifications to identify the key choices that the reported conclusions hinge on. Through the process of narrowing down data analytic specifications, a descriptive critique can systematically alert the scientific community about the gaping holes in the theory, as elements of the theory can be underdeveloped and leave ambiguity in the mapping of the constructs and/or mechanisms.

We further submit that the abductive reasoning in a critique is descriptive when the evaluation scrutinizes whether the statistical results reported in the empirical research that the critique evaluates are accurate by using a set of systematic error-detecting checks [see the "Red Flag" tests recommended and demonstrated by Bergh et al. (2017b)]. A previously published research's descriptive and correlational statistics can be scrutinized in ways such that a data set that is statistically equivalent to the publication's data can be recreated and used to retest the models reported in the publication. The descriptive and correlational statistics would 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). The descriptive mode of evaluation also scrutinizes false positives with simulation-based verification tests that compare reported and expected significance levels (see Goldfarb & King, 2016, for an estimation of how many coefficients reported in a journal may be over- or understated relative to an expected "true" effect size). For the publications that have accurately reported the descriptions of the data and results, the

tests of false positives detect cherry picking of samples or models. Specifically, the tests estimate how likely we would obtain the results that were reported in a publication if the analysis reported in the publication were to be repeated numerous times, with each repetition being done with a new random draw of observations from the same underlying population.

In addition, we submit that the abductive reasoning in a critique is descriptive when the evaluation provides the transparency of the selection between alternative explanations by computing a replication index as they gauge how likely the reported findings may replicate. Deriving such replication index is important, as it can help researchers, reviewers and editors better judge the robustness of the findings. It can also contribute to evaluating whether there is a need for investing time and resource in conducting exact replications. Often time, exact replications are difficult to carry out, especially in field settings, because it is difficult to guarantee the similarity in samples and contexts between the original and the replication studies. To address this issue, Bliese and Wang (2020) proposed to derive *post hoc* statistical power as the replication index either by using the normal distribution or t-distribution of test statistics or by using a nonparametric bootstrap method to resample from the original sample with replacement. The resulted replication index informs how likely the significance of the test statistics for a particular parameter will replicate in future replications.

Finally, we submit that, when the discrepancies are among studies using qualitative and mixed-methods approaches, the mode of evaluation is descriptive when the evaluation discloses the role of researcher cognition, such as transparency of coding process in inductive case research and transparency in analogical reasoning in

interpretive research. Indeed, it is recommended that qualitative research should offer substantive information to convince the readers regarding the transparency of the link between data and empirical generalizations. For example, Golden-Biddle and Locke (2007) argued that one aspect for building researcher credibility was to demonstrate "authenticity" (was the author true to the experience he or she had in the field?).

The Prescriptive Mode of Evaluation

In placing an expectation of compliance to local epistemic values in the selection between alternative explanations, the prescriptive mode of evaluation highlights the preferences of a scientific community. We submit that the abductive reasoning in a critique is prescriptive when the evaluation places an expectation of compliance in regards to increasing the power of a statistical test. For example, as proposed by Simonsohn (2015), if the true effect is zero, a subsequent study needs 2.5 times as many observations as the initial study to have an 80% chance of concluding that the effect is undetectably small. Based on this methodology that can help to illuminate the true effect of interest, Nelson et al. (2018) argue that many of the Reproducibility Project replications (Open Science Collaboration, 2015) were underpowered, with samples smaller than 2.5 times the sample sizes of the initial studies.

We further submit that the abductive reasoning in a critique is prescriptive when the evaluation places an expectation of compliance in regards to using multiple replication studies when examining a specific research finding. For example, the Registered Replication Report (RRR), as a collection of independently conducted, direct replications of an initial study, is a collaborative project that is proposed and approved

before the replication studies are carried out (https://www.psychologicalscience.org/publications/ampps/rrr-guidelines). The journal, Advances in Methods and Practices in Psychological Science, commits to publish a well-prepared RRR report regardless of findings. In facilitating the development of a shared, predetermined protocol, the journal editors contact the initial authors to inform them that an RRR project is in development and to ask for their assistance in providing any materials, code, and data to the authors. When the multiple studies have been completed, a meta-analysis of the results is conducted to assess the size of the hypothesized effect, as well as the degree of effect heterogeneity.

Moreover, we submit that the abductive reasoning in a critique is prescriptive when the evaluation places an expectation of compliance in regards to abandoning dichotomization when assessing the heterogeneity in effect size. This expectation marks a major departure from the conventional NHST paradigm of point estimate (see Gelman, 2018 for problems with the NHST). Dichotomization is rife in the NHST paradigm, causing many problems in social psychology and consumer behavior (McShane & Gal, 2017). A statistical hypothesis is dichotomized as the null versus the alternative; an experimental design is dichotomized as the manipulation being on versus off; an empirical result is dichotomized as statistically significant versus not statistically significant (using a threshold based on p-values, confidence intervals, Bayes factors, or some other purely statistical measure); an interpretation of the result is dichotomized as enough evidence in the data to reject the null hypothesis versus not enough (rejection of the null should not imply acceptance of the alternative); and the conclusion statement is dichotomized as there being ‘an effect’ versus ‘no effect’. Such dichotomization makes

the NHST paradigm unsuitable for quantifying evidence in favor of the null, even when the sample size is large and the p value is close to 1. Shifting toward assessing the heterogeneity 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).

Finally, we submit that, when the discrepancies are among studies using qualitative and mixed-methods approaches, the abductive reasoning of a critique is prescriptive when the evaluation places an expectation of compliance to local epistemic values among qualitative and mixed-methods researchers. As a community, the researchers can emphasize the impartiality of the empirical generalization in inductive case research. They can promote the credibility of analogical reasoning and appropriateness of metaphors in interpretive research. Further qualitative investigation may also be prescribed to generate more comprehensive understanding of the discrepancy with better objectivity and more realism.

Concluding Remarks: Toward Robust and Reliable Knowledge

Increasing concerns about credibility crisis (e.g., Aguinis et al., 2018; Bergh et al., 2017a; Biagioli et al., 2019; Karabag & Berggren, 2012) have motivated many management scholars to provide practical and evidence-based recommendations for research practices of good science (e.g., Banks et al., 2016; Bergh et al., 2017b; Bergh & Oswald, 2020; Bettis et al., 2016a; Bettis et al., 2016b; Bliese & Wang, 2020; Csaszar, 2020; Lee & Wang, 2020; Schwab & Starbuck, 2017; Shaver, 2020; Xu et al., 2020).

A wide range of explanations has been offered in many critiques that seek to diagnose the root causes of the discrepancies between the initial observation and subsequent observations about a theoretical expectation. Having published and reviewed replication studies, we are motivated to understand how a community of scholars in the field of strategic management evaluates the discrepancies.[3] As mentioned earlier, there is not a standardized methodology to guide authors, reviewers, and editors when diagnosing the root causes of discrepancies in cumulative empirical analyses and replication studies. Neither is there an established practice to guide researchers on taking the next step and learning from empirical surprises (Heckman & Singer, 2017: 298).

In the current article, we offer abductive reasoning for evaluating discrepancies in cumulative empirical analyses and replication studies. We propose three modes of evaluation that play different roles for improving the research practices that the scholars in the field of strategic management engage in. We submit that these three modes form the basic evaluation criteria for classifying and telling apart the discrepancies. A normative evaluation calls for pragmatic virtues. A reader can assess a critique's explanations for the discrepancies by asking the following questions: Does the critique identify mechanisms and boundary conditions that may lead to deeper and richer theories? Does the critique recognize that the discrepancies may emerge from the

[3] Lee and Alnahedh (2016) is another example of a replication study on a canonical research question in strategic management. An industry's potential for interdependency among productive activities is one of the central concepts in strategic management. Industry average profitability has been theorized and predicted to peak at moderate levels of interdependency. However, the only empirical test of the prediction (Lenox, Rockart & Lewin, 2006) could not prove that the effect of interdependency on industry average profitability was concave at the standard statistical level of evidence. The lack of empirical support for such an important relationship in our field motivated Lee and Alnahedh to solve the puzzle by conducting a replication and an extension that abducts a mechanism connecting industry characteristics and profitability.

covariation induced by the fact that observations are nested? Does the critique examine the heterogeneity in effect size?

By contrast, a descriptive evaluation calls for the disclosure of cognition in all its idiosyncrasy. A reader can assess the critique's explanations for the discrepancies by asking the following questions: Does the critique compare the measures, manipulations, and exclusions in the publications? Does the critique examine alternative data analytic specifications? Does the critique scrutinize whether the statistical results reported in the publications are accurate by using a set of systematic error-detecting checks? Does the critique compute a replication index as a gauge for how likely the reported findings may replicate?

In comparison, a prescriptive evaluation highlights the preferences of the scientific community. A reader can assess the critique's explanations for the discrepancies by asking the following questions: Does the critique place an expectation of compliance in regards to increasing the power of a statistical test? Does the critique place an expectation of compliance in regards to using multiple replication studies when examining a specific research finding? Does the critique place an expectation of compliance in regards to abandoning dichotomization when assessing the heterogeneity in effect size? Does the critique place an expectation of compliance in regards to adopting registered reporting and results-blind reviewing as alternative publication mechanisms? Collectively, these questions raise sensitivities about quality control for robust and reliable research in a more comprehensive way. These three modes that we propose offer more systematic approaches for authors' sensitivity analysis, journals' peer review process, and editors' guidelines in the evaluation of discrepancies.

The three modes of abductive reasoning in the evaluation of discrepancies signify that, while we embrace robustness and reliability, we accept with humility that we gain knowledge without the certainty we might like. Effects and patterns can and do change over time, and they can look different in different countries and for different groups of people. "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). As we learn to embrace the uncertainty in discovering and building repeatable, cumulative research knowledge, abductive reasoning contributes to the creation of robust and reliable knowledge.

The founding editors of SMR flag "the integration of our research efforts and the construction of a robust, cumulative body of knowledge as key opportunities facing the field" (Leiblein & Reuer, 2020: 2). The current article, which proposes a methodology for building a robust, cumulative body of knowledge, is a provocative essay that responds to the founding editors' statement: "While the negative externalities associated with atheoretical research are not always imposed on researchers (whose careers may actually be advanced by publications of erroneous results), they do affect practitioners who choose to apply practices based on these results" (ibid: 10). We take the statement further by pointing out that erroneous results may disguise as "interesting results." Tihanyi (2020: 329-330) in his editorial commentary warned that, 'Access to large data sets has allowed researchers to find small but interesting effects and to model complex interactions and curvilinear relationships when previously familiar associations "flip" or become the opposite. […] Indeed, authors can falsify their data or p-hack their results because of self-interest or in order to influence societal conversations. […] the process of

trying to engage attention by discovering counterintuitive relationships can easily lead to hypothesizing after the results are known […] The quest for interestingness has resulted in findings that appear only under unique conditions and the construction of complex statistical models that could fall apart in subsequent replication efforts.').

Moreover, we take the founding editors' statement further by complementing the direction of knowledge cumulation with the method of knowledge cumulation. Whereas the direction of knowledge cumulation as championed by SMR points to the canonical research questions of strategic management, the method of knowledge cumulation lacks a standardized methodology to guide authors, reviewers, and editors when encountering discrepancies and learning from empirical surprises. The three modes of evaluation that we propose could help authors in gauging how certain they are about reaching a stopping point for their empirical analysis, aid reviewers in assessing claims of "interesting results," and nudge editors in establishing guidelines for cumulative theoretical development and empirical analyses that are credible and useful to the practice of strategic management.

References

Aguinis, H., Ramani, R. S., & Alabduljader, N. 2018. What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*, 12(1): 83–110.

Aguinis, H., & Solarino, A. M. 2019. Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal* 40(8): 1291–1315.

Anderson, C. J., Bahnik, S., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., . . . Zuni, K (45 co-authors). 2016. Response to comment on "Estimating the reproducibility of psychological science." *Science*, 351(6277): 1037.

Aytug, Z. G., Rothstein, H. R., Zhou, W. & Kern, M. C. 2012. Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods*, *15*(1): 103–133.

Bakker, M., van Dijk, A., & Wicherts, J. M. 2012. The rules of the game called psychological science. *Perspectives on Psychological Science*, 7(6): 543–554.

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.

Berchicci, L., & King, A. 2021. Building knowledge by mapping model uncertainty in six studies of social & financial performance. *Strategic Management Journal*. doi:10.1002/smj.3374

Behfar, K., & Okhuysen, G. A. 2018. Perspective—Discovery within validation logic: deliberately surfacing, complementing, and substituting abductive reasoning in hypothetico-deductive inquiry. *Organization Science*, 29(2): 323–340.

Berger, P. G., & Ofek, E. 1995. Diversification's effect on firm value. *Journal of Financial Economics*, 37(1): 39–65.

Bergh, D. D., & Oswald, F. L. 2020. Fostering robust, reliable, and replicable research at the Journal of Management. *Journal of Management*, 46(7): 1302–1306.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. et al. 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. et al. 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., Ethiraj, S. K., Gambardella, A., Helfat, C.E., & Mitchell, W. et al. 2016a. Creating repeatable cumulative knowledge in strategic management. *Strategic Management Journal*, 37(2): 257–261.

Bettis, R. A., Gambardella, A., Helfat, C., & Mitchell W. 2014. Quantitative empirical analysis in strategic management. *Strategic Management Journal*, 35(7): 949–953.

Bettis, R. A, Helfat, C. E., & Shaver, J. M. et al. 2016b. 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., 2020. Results provide information about cumulative probabilities of finding significance: Let's report this information. *Journal of Management*, 46(7): 1275–1288.

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.
Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280): 1433–1436.
Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9): 637–644.
Chang, S., Kogut, B., & Yang, J. S. 2016. Global diversification discount and its discontents: A bit of self-selection makes a world of difference. *Strategic Management Journal*, 37(11): 2254–2274.
Csaszar, F. A. 2020. *Certum quod factum*: How formal models contribute to the theoretical and empirical robustness of organization theory. *Journal of Management*, 46(7): 1289–1301.
Denis, D. J., Denis, D. K., Yost, K. 2002. Global diversification, industrial diversification, and firm value. *Journal of Finance*, 57(5): 1951–1979.
Ethiraj, S. K., Gambardella, A., & Helfat, C. E. 2016. Replication in strategic management. *Strategic Management Journal*, 11(37): 2191–2192. https://onlinelibrary.wiley.com/toc/10970266/2016/37/11
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.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. 2013. *Bayesian Data Analysis* (3rd ed.), London: Chapman and Hall.
Gelman, A., & Hill, J. 2007. *Data Analysis Using Regression and Multilevel Hierarchical Models* (Vol. 1). New York, NY, USA: Cambridge University Press.
Gelman, A., & Loken, E. 2014. The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American Scientist*, 102(6): 460.
Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. 2016. Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277): 1037–1037.
Golden-Biddle, K., & Locke, K. 2007. *Composing qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.
Goldfarb, B., & Yan, L. 2021. Revisiting Zuckerman's (1999) categorical imperative: An application of epistemic maps for replication. Strategic Management Journal. https://doi.org/10.1002/smj.3290
Goldfarb, B. D., & King, A. A. 2016. Scientific apophenia in strategic management research: Significance tests & mistaken inference. *Strategic Management Journal*, 37: 167–176.

Gupta, V. K., Mortal, S. C., & Guo, X. 2018. Revisiting the gender gap in CEO compensation: Replication and extension of Hill, Upadhyay, and Beekun's (2015) work on CEO gender pay gap. *Strategic Management Journal*, 39(7): 2036–2050.
Heckman, J. J., & Singer, B. 2017. Abducting economics. *American Economic Review: Papers & Proceedings*, 107(5): 298–302.
Hill, A. D., Upadhyay, A. D., & Beekun, R. I. 2015. Do female and ethnically diverse executives endure inequity in the CEO position or do they benefit from their minority status? An empirical examination. *Strategic Management Journal*, 36(8): 1115–1134.
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.
Ketokivi, M., & Mantere, S. 2010. Two strategies for inductive reasoning in organizational research. *Academy of Management Review*, 35(2): 315–333.
King, A.A., Goldfarb, B. and Simcoe, T. 2021. Learning from testimony on quantitative research in management. *Academy of Management Review*, 46(3): 465–488.
Kosová, R., Lafontaine, F., & Perrigot, R. 2013. Organizational form and performance: Evidence from the hotel industry. *Review of Economics and Statistics*, 95(4): 1303–1323.
Leamer, E. E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York, NY: Wiley.
Lee, G. K., & Alnahedh, M. A. 2016. Industries' potential for interdependency and profitability: A panel of 135 industries, 1988–1996. *Strategy Science*, 1(4), 285-308.
Lee, G. K., & Wang, M. 2020. Embracing robustness and reliability in the science of organizations. *Journal of Management*, 46(7): 1238–1243.
Leiblein, M. J., & Reuer, J. J. 2020. Foundations and futures of strategic management. *Strategic Management Review*, 1(1).
Lenox, M. J., Rockart, S. F., & Lewin, A.Y. 2006. Interdependency, competition, and the distribution of firm and industry profits. *Management Science*, 52(5): 757–772.
Locke, K. 2010. Abduction. In A. J. Mills, G. Durepos, & E. Wiebe (Eds.), *Encyclopedia of Case Study Research*, 46–53, Thousand Oaks, CA: Sage.
Mantere, S., & Ketokivi, M. 2013. Reasoning in organization science. *Academy of Management Review, 38*(1): 70–89.
McGrath, J. E. 1981. Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist. 25,* 179–210.
McShane, B. B., & Böckenholt, U. 2018. Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, *83*(1): 255–271.
McShane, B. B., & Gal, D. 2017. Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 112(519): 885–895.
Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. 1999. Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *Lancet*, 354(9193): 1896–1900.
Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251): aac4716.

Peirce, C. S. 1934. Pragmatism and pragmaticism. In C. Hartshorne, & P. Weiss (Eds.), *Collected Papers of Charles Sanders Peirce*, V: 117, Cambridge, MA: Harvard University Press.

Popper, K. R. 1959. *The Logic of Scientific Discovery*. New York: Basic Books.

_____. 1963. *Conjectures and Refutations*. London, U.K.: Routledge and Kegan Paul.

Schmidt, F. L., & Hunter, J. E. 2014. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings* (3rd ed.). Thousand Oaks, CA: Sage.

Schwab, A., & Starbuck, W. H. 2017. A call for openness in research reporting: How to turn covert practices into helpful tools. *Academy of Management Learning & Education*, 16(1): 125–141.

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

_____. 2020. Causal identification through a cumulative body of research in the study of strategy and organizations. *Journal of Management*, 46(7): 1244–1256.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11): 1359–1366.

Simonsohn, U. 2015. Small telescopes: detectability and the evaluation of replication results. *Psychological Science*, 26: 559–569.

Stanley, D. J., & Spence, J. R. 2014. Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, 9(3): 305–318.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5): 702–712.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., & Rennie, D. 2000. Meta-analysis of observational studies in epidemiology. *Journal of American Medical Association*, 283(15): 2008–2012.

Tihanyi, L., 2020. From "that's interesting" to "that's important". *Academy of Management Journal*, 63(2): 329–331.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. 2016. Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113(23): 6454–6459.

Xu, H., Zhang, N., & Zhou, L. 2020. Validity concerns in research using organic data. *Journal of Management*, 46(7): 1257–1274.