LEARNING FROM TESTIMONY ON QUANTITATIVE RESEARCH IN MANAGEMENT

ANDREW KING[*]
Boston University
Questrom School of Business
Rafik B. Hariri Building
595 Commonwealth Ave
Email: aaking@bu.edu

BRENT GOLDFARB
Robert H. Smith School of Business
University of Maryland
College Park, MD 20742
Phone: (202) 730-9484
Email: bgoldfarb@rhsmith.umd.edu

TIMOTHY SIMCOE
Boston University
Questrom School of Business
Rafik B. Hariri Building
595 Commonwealth Ave.
Boston, MA 02215
Phone: (617) 358-5725
Email: tsimcoe@bu.edu

Draft: September 15, 2019

[*] *Corresponding author.* We would like to thank Clark Glymour, Mikko Ketokivi, Edward Leamer, Saku Mantere, Myriam Mariani, PJ Mode, Darrell Rowbottom, Bill Starbuck, members of the University of Oregon (Lundquist) and Rutgers Business School research seminars, participants at the Utah Strategy Conference (2018) and members of the Boston University (Questrom) and University of Maryland (Robert H. Smith School) brownbag research seminars.

LEARNING FROM TESTIMONY ON QUANTITATIVE RESEARCH IN MANAGEMENT

ABSTRACT

Published testimony in management, as in other sciences, includes cases where authors overstate the inferential value of their analysis. Where some scholars have diagnosed a current crisis, we detect an ongoing and universal difficulty: the epistemic problem of learning from testimony. Overcoming this difficulty will require responses suitable to the conditions of management research. To that end, we review the philosophical literature on the epistemology of testimony, which describes the conditions under which common empirical claims provide a basis for knowledge, and we evaluate ways these conditions can be verified. We conclude that in many areas of management research, popular proposals such as pre-registration and replication are unlikely to be effective. We propose revised modes of testimony which could help researchers and readers avoid some barriers to learning from testimony. Finally, we imagine the implications of our analysis for management scholarship and propose how new standards could come about.

(149 words)

INTRO

How do we build our knowledge of business management? Sometimes, we learn by evaluating evidence directly, but more frequently we learn by reading or hearing the testimony of other scholars concerning their own observations and inferences. Yet, most management scholars spend far more time thinking about how they should learn directly from evidence than about how they should learn from the testimony of others. As students, we receive training in the logical requirements for learning from evidence, but almost no instruction in the philosophical issues related to learning from testimony. As researchers, we think about the inferences we make from data, but seldom concern ourselves with the epistemological problems our readers face in trying to learn from our research. In this article, we attempt to rectify this deficiency in our understanding, and by so doing provide guidance on how we should learn from testimony about quantitative empirical research in management.

We believe that our project is both timely and timeless. Several recent studies have argued that reports of empirical research in management may provide an uncertain basis for knowledge. In some circles, this has led to a perception of a field in "crisis", and yet the problem we face is neither recent nor unique (Everett and Earp, 2015, Honig, Lampel et al. 2018). Testimony of all forms has a vulnerability problem. "It is inherent in the nature" of reports from one person to another, Elizabeth Fricker writes, "that insincerity and honest error are both perfectly possible" (Fricker, 1994: p. 145-146). As a result, scholars in all fields, including management, must contemplate how to learn from empirical reports, and they must consider how the best approach for doing so changes with their local empirical context.

To develop a theoretical framework for management scholars and readers, we first review philosophical perspectives on learning from the testimony of others. We assess the conditions under which three common types of empirical claims should be used as a basis for knowledge. We then outline current proposals for assessing the epistemic value of such claims, and we evaluate whether they are likely to be broadly effective in management. Finally, we propose an approach that suits the conditions in many areas of management research. This approach involves the use of what we call *epistemic maps* to connect a "multiverse" of empirical assumptions to inferential outcomes. Such maps, we argue, would encourage authors to engage in better research practices and to report in a more forthright manner. For readers, such maps would allow more direct inference and thereby enable a means of circumventing difficult aspects of testimony's vulnerability problem.

THE EPISTEMOLOGY OF TESTIMONY

As used by philosophers of science, the term "testimony" is defined as *remarks from person A that invite us to accept proposition P* (Coady, 1992: p. 32-33). This definition embraces a broad array of "tellings", which range from driving directions to reports of scientific research (Wilholt, 1985, Fricker, 1995, Wilholt, 2013). Many scholars have noted that testimony is central to our understanding of the world (Coady, 1992, Adler, 2006). For example, few of us have seen red blood cells or evidence of DNA, but we believe they exist. Few management scholars have directly observed the effect of incentives on effort or of training on performance, yet many of us think we know the direction of these relationships, and we "know" these things largely by hearing or reading testimony presented in a variety of settings, such as classrooms, seminars, and publications.

Yet all testimony provides an uncertain basis for knowledge. To use information provided via testimony, the hearer or reader "ascribes to the speaker justification or warrant of knowledge for what she asserts" (Adler, 2006). But what is the basis for granting this warrant? This question represents "the dominant epistemological problem of testimony" (Adler, 2006). One group of scholars, termed reductionists[1], argue that testimony must be validated *a posteriori* before it can be used as a source of knowledge. At the other extreme, anti-reductionists argue that testimony should be trusted "presumptively" – that is without recourse to supporting evidence (Coady, 1992). In between these two positions, philosophers such as Elizabeth Fricker argue for a contingent approach to testimony based on local conditions (Fricker, 2004).

The main premise of the reductionist thesis is straightforward: testimony, because it relies on another human, is vulnerable to error and insincerity. As a result, reductionists argue that the use of testimony can only be justified via recourse to other sources, such as direct perception or prior experience. The Scottish philosopher David Hume is typically accepted as a proponent of the reductionist perspective (Adler, 2006). He argues that testimony should be trusted only when it conforms with "reality", or when its source has been shown to be reliable. Since Hume, the reductionist position has been repeatedly attacked. As a result, an anti-reductionist position is now ascendant among philosophers of science (Adler, 2006).

Anti-reductionists argue that *a posteriori* evaluation of testimony is seldom feasible, so we should therefore trust testimony "presumptively". For example, CAJ Coady (1992) argues that 1) our beliefs are largely based on testimony that we have trusted without evidence; 2) these beliefs are evidently helpful; and critically, 3) there is rarely a feasible way to assess the testimony we receive. As a result, Coady concludes that granting testimony a presumptive "warrant of authority" is "the only honest [position] to adopt" (Coady, 1992: p. vii).

[1] "The term 'reduction' as used in philosophy expresses the idea that if an entity x reduces to an entity y then y is in a sense prior to x, is more basic than x" (Van Riel and Van Gulick, 2016). With respect to theory development, reductionists are those that think science can and should advance by substituting new and more general theories for older more specific ones. With respect to the use of testimony, reductionists believe that direct experience is prior and more basic than testimony. The two uses and groups are not identical.

The philosopher Elizabeth Fricker (1994) proposes a middle ground between the two sides. She decries Coady's recommendation of presumptive trust as "an epistemic charter for the gullible and undiscriminating" (p. 126), but she also admits that for many types of testimony, it is simply too difficult to check its veracity. Users of testimony, she argues, should consider local conditions before deciding how to approach testimony, because in some circumstances they will be able to ascertain whether that testimony meets the necessary conditions for "veridicality."

Fricker (2004) points out that hearing or reading testimony, like any experience, provides a basis for knowledge when its operation is veridical – that is the experience links to something real. For example, the operation "seeing" is veridical if the viewer's perception of an object corresponds with an actual object. Similarly, the operation "read testimony" is veridical if the author is competent to advance her proposition and expresses it in a sincere way. For empirical research, this means that an author's claims are justified by proper empirical analysis and reported in a forthright way. Readers or hearers cannot determine if claims are *true* or *false*, Fricker (2002) reasons, but they can evaluate whether claims satisfy veridicality conditions. She contends that by reframing testimony in this way, readers can decide how to use a research report, but she does not provide instructions for how they should do so (Fricker, 2002). In this paper, we take up the challenge of proposing an approach for evaluating veridicality that is suitable to conditions common in management research.

The challenge we set is daunting, perhaps even insurmountable. Many authorities on the sociology of science are less sanguine than Fricker about the opportunity for local evaluation of the veridicality of empirical reports. Helen Longino (2002), for example, argues that a reader's ability to evaluate an empirical claim is severely limited by the incompleteness or imprecision of published reports. She notes that empirical science requires countless methodological decisions that are “underdetermined” by the research questions or data. Gelman and Loken (2016) illustrate the importance of this point by comparing the empirical process to a stroll through a “garden of forking paths.” At each fork, the empirical researcher makes a choice that influences where they exit the garden – that is the estimates obtained and inferences formed. These choices, or the logic behind them, are seldom fully documented (Douglas, 2000). Such incomplete reporting, Torsten Wilholt (2013) argues, prevents readers, even expert ones, from evaluating whether the researcher had justification for the claims she advanced. Assessing the veridicality of claims, he argues, would require peering into the mind of the researcher to understand her assumptions, values, and goals. Only then could the reader fill in missing information about what empirical choices led to the reported claims. Wilholt admits, however, that his analysis is limited to particular types of empirical claims; he does not consider how his assessment might be effected by “local” norms of analysis and reporting.

In fact, none of the above assessments are situated in the management literature and thus do not enumerate or evaluate the empirical claims typically made by authors in our field. To assess the feasibility of Fricker’s local reduction, we must first consider the kinds of claims common in management. Doing so will allow us to understand their veridicality conditions, and inform our analysis of whether or when reduction is possible.

CONDITIONS FOR VERIDICAL RESEARCH CLAIMS

In this section, we consider common types of empirical claims advanced in the management literature, and evaluate their veridicality conditions: that is, when authors are competent and sincere in advancing them. For scientific claims, competence means that the speaker has a basis that is justified by the epistemology of science (Fricker, 2002). For quantitative empirical analysis, this means that the scientist must somehow circumvent David Hume’s argument that no claim to knowledge is *ever* warranted. All inferences from evidence, Hume argued, require supporting assumptions, and since these assumptions cannot be validated independently, all inferences are vulnerable to error (Hume, 1748/1993). This "enormously important result" means that Hume has since "loomed like a colossus" over the development of the philosophy of science (Stove, 1982: p. 72). Many scholars have attempted direct assaults on his argument, but these have largely fallen from favor and been abandoned (Hacking, 2001). Approaches that avoid, rather than overcome, Hume's problem of induction have been more successful, and provide justification for the types of claims commonly expressed in the management literature.

Although the precise language used to describe "key findings" in the management literature varies from one paper to the next, authors usually advance claims that fall into one of three broad categories: *explanations, frequencies,* or *beliefs*.[2] Each type of claim differs in strength and nature, and each comes from a different tradition in epistemology. These traditions both enable and constrain what scholars can infer from evidence. They provide a means for circumventing David Hume's argument that *no* claims to knowledge are ever warranted; yet they constrain authors by requiring them to follow exacting empirical procedures and to express their inferences in a specific manner.

Competence to make common claims

An *explanation* is a conjecture about an observed pattern of evidence. Explanations are often presented in management reports as part of "post hoc analyses of alternative patterns in data, and in discussion sections where non-significant or unanticipated results are speculated upon or

[2] The latter two have also been called "frequency and belief probabilities" (Ian Hacking) "probability$_2$ and probability$_1$" (Rudolf Carnap), or aleatory and epistemic probabilities (JM Keynes).

where links to other findings are proposed, [and] where mysteries are explored" (Behfar and Okhuysen, 2018: p 327).[3] An example of an explanation drawn from the management literature is found in Zaheer and Soda (2009: p. 28): "A possible explanation for this result is that we investigate the redundancy of the network at the team level of analysis, where factors such as efficiency and routines…may be exerting stronger influences on performance."

The philosopher Charles Sanders Pierce was an early student of such explanations, and he termed the process of finding them "abduction", from the Latin "to take away". Pierce hoped to develop a basis by which abduction could lead to truth claims, but he eventually concluded that abduction provided justification for nothing more than "guesses." In the past few decades, some scholars have argued that the epistemic virtues (e.g. likeliness, simplicity, elegance, fruitfulness, etc.) may provide a basis for selecting some explanations over others (Lipton, 2003, Ketokivi and Mantere, 2010). At present, however, most epistemologists agree that abduction provides justification only for conjectures about possible explanations for observed evidence (Adler, 2006). For example, Schurz (2008: p. 205) argues that abduction allows only a basis for making a "promising explanatory conjecture" which then must be "subject to further test." Framed this way, *explanations* avoid Hume's skepticism with respect to induction by making no claim that knowledge has been inferred from evidence. In accepting *explanations*, readers do not need to face Wilholt's challenge that they must perceive and agree with researcher choices and values to determine if a claim is veridical. The reader need only conclude that the evidence exists and the explanation is plausible. In the management literature, explanations are frequently debated and discussed, but are accepted as veridical so long as they have an evidentiary basis and advance no stronger claim (Behfar and Okhuysen, 2018).

[3] They use the term "plausible knowledge claim", but we prefer the more standard "explanation".

Frequency Claims are also common in the management literature. Chatterjee and Hambrick (2007: p. 216) provide an example of a classic frequency claim: “Both measures of objective performance were significantly positively associated with risky outlays… at $p < .05$ and …at $p < .01$.” Frequency claims can also be expressed in terms of a “confidence interval”, that is a prediction for how often an estimate would fall within a specified range. The statistical logic of these approaches was developed, respectively, by Ronald Fisher, and Jerzy Neyman/Egon Pearson (Schneider, 2015). Both methods provide a means of avoiding Hume’s problem of induction by allowing inference to rational action without the need for interpretation with respect to the truth (Hacking, 2001). For example, a clinical trial of a drug may estimate the conditional frequency of positive or negative outcomes, and thereby allow actors to make rational cost/benefit decisions, without ever needing to know whether the drug itself caused those outcomes.

Frequentist analysis places strict limitations on the empirical process researchers must follow, because frequency claims describe what is expected to happen if an *identical* test were conducted on an *identically* constructed sample from the same population. Practically, this means that authors must specify in advance (a) the hypotheses to be tested, (b) the sampling plan and population to be used,[4] (c) the parameters to be evaluated, and (d) the test statistics to be employed (Spanos, 2010). Even slight freedom to deviate from these plans, will undermine frequency claims (Sanborn and Hills, 2014). For example, if a researcher does not set a test plan in advance, she may not know how to interpret multiple tests (Fisher, 1960). Should she, for example, consider a particular test statistic to be a) independent or b) part of a connected family of tests? If she has not

[4] The ability to stop or continue gathering data (what epistemologists refer to as “optional stopping”) can undermine frequency claims, even when the decision to continue or stop does not depend on the data being gathered or the estimate obtained (Mayo, 1996).

specified her rule in advance, she is unlikely to know how they should be interpreted, and as a result, she is not competent to make a precise frequency claim.

The importance of pre-specification creates a problem for readers who encounter frequency claims in management research. A reader can inspect the statistical process and analytical methods that were used, but she cannot ascertain when or how that process was chosen. She can evaluate whether an appropriate statistical method was employed, but this provides only a *necessary (and not sufficient)* condition for justifying frequency claims. To know that the author is competent to advance a frequency claim, the reader must know that empirical choices were set in advance. Thus, consistent with Wilholt's (2013) argument, to establish veridicality the reader or hearer of an empirical claim must somehow see into the mind of the researcher to ascertain when and why they selected a particular research design.

Several scholars in management have pointed out that conditions for veridical frequency claims are commonly violated. For example, Richard Bettis (2012) has reported that, based on his personal experience, scholars often search through data for plausible hypotheses; and prominent scholars have admitted to following and advocating this practice (Bartlett, 2017). The problem is not limited to management. In psychology, John, Loewenstein, & Prelec (2012) & report that a majority of the researchers they surveyed admitted to adjusting their sampling plan during their analysis. In economics, Nobel Laureate James Heckman and statistician Burton Singer observe that most scholars violate the conditions for justified frequentist claims: "Peeking at the data and formulating and building new models in the light of those views is an often-committed frequentist sin" (Heckman and Singer, 2017, p. 299).

Belief claims express how hypotheses should be understood in light of observed evidence, i.e. the probability that a hypothesis is true conditional on the evidence, P(H|E). Such claims often represent a researcher's ultimate summary of their analysis. A good example of a belief claim can be found in Sine and Lee (2009, p. 151): "Our analysis indicates that the presence of local social movements was responsible for this regional variation [in wind energy entrepreneurship]." Because belief claims necessitate extrapolation from the physical realm of experience to the epistemic realm of truth, they must surmount Hume's objection that such induction is never justified. How can this be overcome? The most enduring approaches have sought to avoid the problem by shifting the nature of belief claims. The goal is to allow a rational *approach* to knowledge without ever claiming confirmation of knowledge itself (Hacking, 2001). Scholars working within the traditions of logical probability and classical statistics have both proposed justifying conditions.

One approach, loosely termed Bayesian, concedes that Hume is correct that we are never justified to say we know the truth, but contends that we can know "whether we are reasonable in modifying [our] opinions in the light of new experience" (Hacking, 2001: p. 256).[5] Unfortunately, a fully Bayesian approach to learning faces both logical and practical difficulties, and thus is rarely used in many areas of social science, including management (Glymour, 1980; Senn, 2011). To rescue this approach to justification of belief claims, Hacking (1965) argues that most scholars can and do justify *directional* arguments about belief using a "law of likelihoods". This law is easily derived from Bayes Rule, yet it obviates the need for scientists to specify ex-ante probability distributions. It also provides possible justification for commonly made claims that evidence "supports" a proposed hypothesis. A common form of the law states that: if the evidence (E) is

[5] This tactic has the added benefit that it avoids some classic problems of induction (e.g. black swans) and accommodates some commonly practiced methods in science, such as falsification.

more likely conditional upon a preferred hypothesis (H) than it is conditional on all alternative hypotheses (!H), then the evidence supports belief in H.[6]

Given the widespread use of non-Bayesian methods, some scholars have sought to develop a justification for belief claims that relies only on classical statistics (Mayo, 1996, Mayo and Cox, 2006, Mayo and Spanos, 2011). Though their program is ongoing, their initial statements about justifying conditions closely mirror the law of likelihoods. Researchers are justified in claiming that the evidence provides "support" for a hypothesis, they reason, when the data agree with the preferred hypothesis, *and* there is a low probability that the data agree with *any* alternative hypothesis. Thus, for both Bayesian and non-Bayesian epistemologists, to make a belief claim it is necessary *but not sufficient* to provide evidence that is consistent with a preferred hypothesis. For a belief claim to be veridical, the author must also rule out alternative mechanisms and rival hypotheses that could generate the same result.

Because belief claims are only justified if an author considers alternative explanations for the evidence, they are closely linked to what econometricians call the *identification* problem.[7] Randomized controlled experiments provide one powerful method for ruling out rival hypotheses, but other modes of analysis are also commonly used. Indeed, James Heckman (2000) argues that a major contribution of 20th century econometrics was the development of methods for demonstrating that a particular explanation is identified, in the sense that a statistical association can be connected to a particular hypothesis. The resulting "identification revolution" has had a strong influence on empirical research in economics and management (Angrist and Pischke, 2010).

[6] This law can be stated in terms of the likelihood ratio $P(E|H)/P(E|!H) > 1$, where $P(E|H)$ is the conditional probability of observing evidence E given the truth of hypothesis H.

[7] Manski (1995) defines identification as "seek[ing] to characterize the conclusions that could be drawn if one could use the sampling process to obtain an unlimited number of observations" and notes that, "identification problems [pose] inferential difficulties [that] can be alleviated only by invoking stronger assumptions or initiating new sampling processes that yield different kinds of data."

Distinguished scholars have noted that confusion about the necessary conditions for making a belief claim often leads to erroneous claims (Schwab, et. al, 2011). For example, Schneider (2015: p. 421) argues that many scholars "regard p values as a statement about the probability of a null hypothesis being true or conversely, $1-p$ as the probability of the alternative hypothesis being true". Other scholars have identified cases where researchers simply fail to consider important rival explanations, such as endogeneity or mediation, and move directly from evidence that is *consistent* with a hypothesis to a claim that the evidence *supports* that hypothesis (Antonakis, et. al, 2010; Nickerson and Hamilton, 2003; Shaver, 2005). In the above cases, *necessary* justification for belief claims can be ruled out relatively easily.

Readers face a greater challenge in determining that researchers had *sufficient* justification for the belief claims they advanced. First, they must evaluate whether all reasonable alternative explanations were considered. If not, the law of likelihoods will deliver a misleading result: it will imply belief in a particular hypothesis, when in fact this hypothesis is simply the "best of a bad lot" (Douven, 2002). Second, the reader must ascertain the assumptions that enabled identification of the connection between the evidence and particular hypotheses (Leamer, 2010). Because belief claims are always contingent on acceptance of a set of maintained assumptions (Manski, 1995) readers hoping to use those claims as a basis for knowledge must be able to observe, and endorse, those assumptions. Thus, once again, for belief claims, we encounter Wilholt's barrier to evaluating the veridicality of testimony: given the incompleteness of empirical reports and our inability to observe researcher assumptions, we cannot reduce testimony about belief claims.

Sincerity

Above we discuss the conditions under which a speaker or writer is *competent* to advance an inference claim. For such testimony to be veridical, a competent speaker must also express the claim in a "*sincere*" way (Fricker, 2004).

As used by philosophers of science, "sincerity" has a deeper meaning than honesty or conformance with an obligation to "say what is true". Sincerity requires the speaker to make claims that accurately reflect their understanding (Elgin, 2002). For example, it is insincere, although true, to say that "Beth is somewhere in the United States" if one knows she is in Minneapolis. Likewise, a sincere author cannot report only those statistical models with stronger (or weaker) results, even if those results were properly calculated, unless they have a reason to believe that such results are most informative. Sincerity also includes a requirement to provide information in a way that will convey the proper understanding of the speaker's knowledge. Within management science, this requirement to consider the reader means that reports must reflect norms of communication. For example, researchers can not imply doubt or confidence that they do not actually feel.[8]

Scholars in all fields have noted misrepresentations of inference claims. Most commonly, estimates obtained through an abductive search for an *explanation* are represented as a valid *frequency claim* (Bettis, 2012). The practice of HARKing, for Hypothesizing after Results are Known, is a common example of such misreporting. In HARKing, the researcher evaluates many possible combinations of explanatory and outcome variables. After finding a relationship with attractive properties, the researcher develops an explanation for this relationship. So far, the

[8] In the case of published testimony, the requirement for sincerity is shared by the reviewers and editors of the journal. Sincerity means that the published testimony must not misinform the journal's readership. Notably, the common practice of selecting publications based on the strength or significance of a result is a clear violation of sincerity – whether this selection occurs at the level of the researcher or that of the journal.

researcher is following a valid abductive process and is competent to claim a possible explanation for the observed evidence. However, many researchers report as if they had followed a frequentist test process, make frequency claims, and thereby encourage readers to accept a stronger inference than they are justified to make.

John et al. (2012) find evidence that many scholars engage in reporting practices that reflect their knowledge in a biased way. For example, more than half of those surveyed admitted to failing to report all of the dependent variables that had been tested, and 45% admitted to selectively reporting studies that "worked." Such selective reporting has been noted in many areas of social science. In his article, "The Search for Asterisks", Richard Bettis (2012) describes practices whereby researchers hunt for interesting explanations and then misrepresent these as the product of frequentist tests. In strategy, Goldfarb and King (2015) report that the distribution of test statistics from 300 articles implies a reporting process that selects for statistical significance.

With respect to misrepresentation of belief claims, prominent scholars argue that authors commonly fail to represent the plausibility of rival explanations, and thus mislead readers about the strength and justification of claims (Antonakis, et. al, 2010; Nickerson and Hamilton, 2003). Even when authors provide identification arguments, they may not highlight all of the underlying assumptions, or provide any evidence that those assumptions are plausible (Angrist, 2009). Scholars may also try multiple strategies for identification, and then pick the one that contains a desired collection of results (Leamer, 2010).

Readers often face great difficulty in determining the "sincerity" (in the philosophical sense of fully authentic and neutral reporting) of authors because they cannot observe the set of estimates from which the reported one was selected. This has led some authors to adopt a skeptical approach to reported frequency claims. All of the concepts of frequency analysis, Ed Leamer (1983: p. 37) writes, “utterly lose their meaning by the time the researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose.” Heckman and Singer (2018) concur, noting that “[t]est statistics are reported as if the hypotheses being tested did not originate from the data being assessed.” Such skepticism also extends to reporting of belief claims: “Most authors”, Leamer (2010: 36) writes, “leave the rest of us wondering how hard they had to work to find their favorite outcomes and how sure we have to be about the instrumental variables assumptions [they made] with accidentally randomized treatments…It's like a court of law in which we hear only the experts on the plaintiff's side.”

PROPOSALS FOR ALLOWING REDUCTION OF EMPIRICAL TESTIMONY

If reports of quantitative research include claims that are not veridical, how should scholars learn from the literature on quantitative empirical research? Scholars in management and its abutting disciplines have proposed approaches for improving the ability of readers to reduce empirical claims. The first is to try to reduce published testimony by evaluating the author’s ability to deviate from a justified empirical method; and the second is to reduce published testimony by evaluating, usually through replication, correspondence among findings. These approaches have become so well-known that their effectiveness is often taken for granted. In fact, as we will demonstrate below, leading epistemologists and practicing empirical scholars are skeptical that either approach will be effective or practical across broad areas of social science, including management. As a result, we contend, the opportunity for Fricker’s “local reduction” may be local indeed, and readers may often need to employ a default rule – either trust or doubt – when evaluating testimony. Since neither trust nor skepticism seems to provide a good way forward, we propose that an alternative approach to testimony is needed – one that helps researchers and readers circumvent the most difficult aspects of assessing the veridicality of testimony.

Reduction through Evaluation of Researcher Constraints

One common proposal for assessing the justification for frequency claims involves the evaluation of constraints on researcher choice (Fricker, 2002). Sometimes these constraints are exogenously imposed. In the best of circumstances, the precedence of a previous study may require adherence to specific hypotheses, data sources, sampling plans, test plans, and analyses. To the extent that such constraints makes empirical assumptions visible and verifiable, it may help readers to evaluate the justification of replication research. More generally, empirical standards, enforced by academic communities, may limit author choice. But such standards are usually insufficient, Wilholt (2013) argues, because "[t]he ever-developing practices of science will always force their practitioners to make choices… that cannot be constrained by conventional standards" (2013: p. 244-245).

When external constraints are weak or absent, researchers may choose to specify their plans in advance, register them with a third party, and thereby make their empirical process more visible and their claims more credible. This approach has recently gained support from many journals and academic organizations. The hope is that researchers who pre-register can credibly profess to have followed an approach that allows justified claims (Kidwell, Lazarević et al., 2016). Pre-registration can also limit the problem of selective reporting by helping scholars observe whether reported claims have been selected from a large pool of attempts.

For research reports that eschew frequency claims and make only belief claims, strict pre-specification is not required. A researcher operating within this paradigm can try multiple specifications and identification strategies and yet still be competent to make a justified belief claim. But belief claims also create a new set of difficulties, because readers often cannot inspect the identifying assumptions or observe the probability of the evidence conditional on rival explanations.[9] To solve this problem, scholars have proposed that researchers limit themselves to empirical methods that possess a kind of inalienable credibility; that is, methods that require few assumptions and also circumvent the problem of alternative explanations. For example, Angrist & Pischke (2010: p. 4) argue that "design-based studies", which typically rely on natural experiments, "are distinguished by their prima facie credibility."[10] Other scholars go further. For example, statistician Donald Rubin argues that only randomized control trials allow causal belief claims. He describes claims based on other approaches, such as tests of possible causes of an observed effect, "as more of a cocktail conversation topic than a scientific inquiry" (Li and Mealli, 2014: p. 446).

Unfortunately, constraints on researcher choice, whether through pre-registration or limitations on empirical design, can be costly and ineffective. Indeed, one of the most vocal advocates for more credible research, statistician Andrew Gelman (2014), rejects pre-registration as a viable way forward. Practical applications of preregistration, such as ex-ante specification of hypotheses, are insufficient, he argues, because they still leave researchers with "too many data-processing and data-analysis choices". On the other hand, Gelman contends, "'full preregistration' or a 'complete step-by-step plan'…does not work in areas where data must be explored before they can be understood and used." Social science research, he argues, is often conducted on inchoate questions using uncharted data. "Many of my most important applied results were

[9] When the evidence for belief claims comes from interpretation of frequentist statistics, such as p-values, the preceding difficulties with pre-registration and replication continue to apply.

[10] Because all design-based studies rely on maintained assumptions, some scholars have proposed that the "etiquette" of reporting should be increased to allow readers to evaluate those assumptions. For a discussion of supplemental tests see Athey and Imbens (2019), and a proposed "etiquette" for incorporating these ideas can be found at http://people.bu.edu/tsimcoe/etiquette.html.

interactions that my colleagues and I noticed only after spending a lot of time with our data" (Gelman, 2014). Heckman and Singer (2017) agree with Gelman that pre-specification will impoverish researcher learning without providing compensating gains in credibility.

Several leading scholars have also expressed skepticism that constraints on empirical design, including the use of randomized control trials (RCT), will effectively assure the credibility of empirical claims (Heckman, 2008; Leamer, 2010, Simmons et al.,2011). Simmons, Nelson, and Simonsohn (2011) provide a stunning example of how an RCT can be manipulated to generate a seemingly credible, but false, empirical "result." By searching through a number of possible variables, they find "significant" evidence that listening to the Beatle's song, "When I'm 64" (rather than "Kalimba") makes people younger. Leamer (2010) also catalogs a number of ways that scholars can manipulate results from randomized control trials.

Several scholars point out that the cost of limiting empirical designs may exceed any entailed gains in credibility (Heckman, 2008; Gelman & Imbens, 2013). Heckman (2008) notes that randomized experiments circumscribe what authors can investigate. Many important questions in social science, he argues, are not amenable to researcher manipulation, and thus researchers must rely on theory to provide identifying assumptions, even if doing so reduces the inherent credibility of their reports. Thus a tradeoff exists, Heckman contends, between credibility and empirical reach. In a recent publication, Banerjee, Chassang, and Snowberg (2017) formalize this idea of a tradeoff between learning and credibility to evaluate how researchers may alter research designs when faced with varying testimonial barriers.

In summary, some of the top empirical social scientists of our era, Leamer, Heckman, Singer, and Gelman, are all skeptical of the use of constraints to allow the reduction of testimony.

Reduction through Evaluation of Correspondence with Other Research

Traditionally, scholars have proposed that the veridicality of testimony can be determined, *a posteriori*, by correspondence with other information. For example, David Hume argued for assessment of testimony based on correspondence with other evidence, and several modern proposals for improving empirical research follow this approach. Stefan Schmidt (2009) argues that systematic replication can uncover fraud and enable verification of empirical claims.

Conceptually, the logic of replication is straightforward. Any result from a single sample represents merely one draw from a distribution of probable results, and thus provides an inexact basis for inference. If hidden search processes produce any sort of bias, then effect sizes, frequency estimates, and posterior probabilities will all be distorted. Further draws from the same population (e.g. replication), and unbiased calculation of estimates, are required for convergence to true values.

Unfortunately, many practical barriers exist to replication. For example, the use of replication studies can be impeded by selective and incomplete reporting. As a consequence, readers (and analysts) may only observe those studies with results that researchers or journals thought desirable (Tatsioni, Bonitsis et al., 2007; Rosenthal, 1979; Simonsohn, Nelson et al., 2014). If this "file drawer problem" is not solved, replication analysis will deliver faulty conclusions.

To solve problems caused by one mode of selection (on significance), Simonsohn, Nelson, and Simmons (2014) have developed a meta-analytical approach that assumes the publication process reveals only those studies with "significant" results (e.g. $p<0.05$). They show that readers can still use the distribution of reported test statistics to evaluate the veracity of the *entire* collection. Selection of an alternative type can create other difficulties in assessing replications.

If authors or journals select publications based on surprising or "interesting" results (Davis, 1971), then authors may generate incoherent findings (Young, Ioannidis et al., 2008). De Long and Lang (1992) propose that just such a mechanism has occurred in economics. They reason that authors and journals tend to report "surprising" results, and this means that these findings are usually wrong. This leads De Long and Lang (1992: p 1259) to pose "a very peculiar epistemological problem": How can a "rational reader" learn from the literature if each published paper should be interpreted as demonstrating the opposite of what is claimed?

The problem of selective reporting is made graver by an overall lack of replication studies. Helen Longino (2002) notes the existence of "a gap between the ideal of replication … and the reality." Even in medicine, where replication seems literally an issue of life and death, few studies are repeated (Ioannidis, 2005). In management, replications are even less common. The incoming editor of the Academy of Management Journal, Laszlo Tihanyi (2019), conducted an assessment of ten years of publications in the journal, and only uncovered ten replications, or about 1% of the total. This situation is improving, but barriers to replications still exist. In some areas of management, scholars use private data to investigate the effect of unrepeatable interventions. Each of these represents a unique case, studied with a particular sample, from a finite population; thus, no exact replication can ever be conducted. In some areas, replications are still regarded as minor contributions, and thus avoided by ambitious scholars. Summarizing conditions common across many fields, Longino (2002) concludes that the "widespread" scarcity of replication has made it difficult to evaluate the veracity of published work and thereby contributed to a perception of a crisis in credibility (Longino, 2002).

If the scarcity of exact replications means that they are unlikely to provide an avenue for assessing the credibility of frequency or belief claims, perhaps such claims can be checked ex-post by coherence with more distantly related knowledge, such as conceptually connected empirical reports (Psillos, 2002, Mackonis, 2013). In psychology, for example, Schmidt (2009) argues that conceptual replications can be used to validate underlying hypotheses. In a conceptual replication, similar theories are tested in new settings. Unfortunately, this generally implies that authors must use new measures and sampling schemes that make sense in their new research setting; and aggregating and comparing results across these different settings is often difficult or infeasible (Heckman, 2008, Leamer, 2010).

Within management, scholars have also argued that belief claims should be judged based on the degree to which they cohere with existing theory. For example, Lounsbury and Beckman (2015 suggest that readers should not update beliefs if the [organizational] theoretical reasoning used in empirical reports seems to be ad-hoc. Consistently, Heckman (2008) warns that aggregating and evaluating evidence from conceptually related studies requires that researchers and readers all share the same body of formal theory. In our view, the cross-disciplinary nature of management research makes reliance on coherence an unlikely tool for generating consensus among management scholars.

In summary, correspondence across research reports is unlikely, for many areas of investigation, to allow a means for readers to assess the veridicality of empirical claims.

ALLEVIATING THE PROBLEM OF TESTIMONY IN MANAGEMENT

We have argued that scholars in management face a dilemma with respect to the use of empirical testimony as a basis for knowledge. There is ample evidence that many reports are not reliable sources of credible empirical claims, and reducing such claims by determining their veridicality appears infeasible. Commonly proposed solutions for enabling the reduction of empirical claims, such as constraints on empirical choice or ex-post validation through replication, are likely to be ineffective and impractical for many types of research. What, then, is the right way forward for management scholars?

One approach to the vulnerability problem of testimony is for readers and researchers to retreat to claims for which veridicality can be more easily assessed – that is explanations. Leamer (2008: pg. 3) argues that readers should treat all empirical claims as explanations. “The words ‘theory and evidence’ suggest and incessant march toward a level of scientific certitude”, he argues, “the words ‘patterns and stories’ much more accurately convey our level of knowledge, now, and in the future as well. It is literature, not science”. James Heckman and Burton Singer (2008) agree, “Economists should abduct” and present their claims accordingly, as plausible explanations of observed evidence. Consistently, a prominent group of empirical scholars have proposed abandoning frequentist tests and certain related claims (McShane, Gal et al., 2017). In 2016, the editors at the *Strategic Management Journal* banned reporting of critical values for frequentist test statistics (e.g., p < .05 or .01).

We agree that research that uses abduction to develop plausible explanations is well suited to the management research setting. Abduction fits the types of data that are normally available and the limitations that readers face in assessing stronger claims. Pre-specification is impractical for most research conducted on archival datasets, and replication is rare. Readers cannot observe the mental state of the researcher as they maneuvered through the “forking paths” of choices that are almost inevitable in empirical research that seeks to justify frequency or belief claims. Nor can they simply assume that they share with the researcher a set of deep values, assumptions, and goals that enable them to assume the researcher has done what they would have done in her place (Wilholt, 2014: pg. 248). To us, all these considerations recommend that the “coordination problem” between researchers and readers could be resolved by the use of abduction and the forthright reporting of explanations.

Nevertheless, limiting our aspirations to the identification of plausible explanations seems excessively pesimistic. To build a solid body of cumulative knowledge, we must go beyond plausible conjectures. Although we support abductive research, we also believe that it is possible, in some cases, to overcome the vulnerability problem of testimony with respect to stronger claims. As discussed earlier, a principal difficulty in assessing frequency or belief claims is that the reader usually cannot observe the competence of the researcher to make those claims. Central to this difficulty is the need for the reader to peer into the researcher's brain to determine her assumptions, values, and goals – and then estimate their impact on her empirical choices. But what if the roles of the researcher and the reader could be adjusted to allow the reader to consider her own mental state and not that of the researcher? The reader could then make inferences more directly from data and avoid the need to consider the mental state of the researcher. Readers would still need to consider the researcher's sincerity, but some of the most difficult aspects of verifying competence could be avoided.

The logic of the idea can be understood by analogy to the use of a road map. If a lost driver asks a stranger for the best way to reach a destination, a problem of testimony is encountered. The hearer cannot observe the stranger's mental state to confirm that they know the best route, nor can she determine whether the route chosen by the speaker would match the one she would judge best. However, if the stranger pulls out a road map and hands it to the driver, her problem is much reduced. The responsibility for selecting a route has now shifted from the stranger to the driver. Of course, she still must trust that the map is accurate and unbiased, but conditional on this assumption, she can consider her *own* competence in selecting a route. She can determine if a particular path reaches a particular destination, and she can consider her own assumptions (traffic, comfort, etc.) in making her choice. In the case of testimony about quantitative empirical research, an analogous "road map" would show the connection between a large number of feasible assumptions and estimates. For quantitative research in psychology, Steegen et al. (2016) propose the use of "Multiverse Analysis" to improve research transparency. They suggest that "instead of performing only one analysis, researchers could perform …all analyses across … a large set of reasonable scenarios. (Steegen, Tuerlinckx et al., 2016: p. 702)"

A map of a "multiverse" of analyses, or what we term an "epistemic map[11]", connects assumptions to estimates, and thereby allows a reader to form her own inferences based on her own assumptions. If the reader is uncertain about her preferences, she can inspect the broad pattern of results. If she has a strong prior belief in particular assumptions, she can use the map as a kind of inference look-up table – indexing from her assumptions to the resulting estimates. In this new division of labor, the researcher's role has changed from claims-maker to cartographer. "The job of a researcher", Leamer (1983: p. 38) argues, "is then to report economically and informatively the mapping from assumptions into inferences. In a slogan, 'The mapping is the message.'" Similar arguments have been advanced by other prominent scholars. In psychology, Simonsohn, Simmons, and Nelson (2015) have proposed that researchers map multiple possible specifications on what they call a "specification curve". In economics, Heckman and Singer (2017:p. 301) advocate more exhaustive disclosure of multiple analyses, and they hope that doing so "encourages readers of such studies to form their own opinions."

Mapping liberates researchers to conduct and communicate a broad range of empirical claims. Researchers can use an abductive process to search through multiple specifications and propose the best ones, so long as they map the options they considered. Alternatively, a researcher who conducts

[11] This term seems to us to capture the use of the map to evaluate "knowledge or its degree of validation" – the definition of epistemic.

a true empirical test can advance a stronger claim (frequency or belief) and use the map as a demonstration of their claims robustness to different assumptions. Returning to our analogy, they researcher can communicate "this is the route I would take (expressing whatever confidence is appropriate), but feel free to decide for yourself." As with geographical maps, epistemic maps can be enhanced by including a "legend" that provides additional information. For example, the legend to an epistemic map could indicate which analyses correspond to the authors' preferred specification, or which analyses conform to the assumptions used in prior research. When creating an epistemic map for use in helping readers form inferences to causation, the legend could provide information about the maintained assumptions associated with various identification strategies, and the extent to which those assumptions have been subjected to empirical scrutiny.

Regardless of the claims advanced by researchers, epistemic maps enable readers greater ability to make their own inferences. She can envision a particular test plan, determine if it has been mapped; observe the resulting estimates, and evaluate its robustness to alternative assumptions. This ability to index from particular assumptions to estimates should, in principle, allow a user to conduct her own frequentist test. Epistemic maps can also facilitate belief claims by bounding the effect of rival explanations. For example, a researcher could identify alternative theories and propose "identifying strategies" for limiting the effect of these rival explanations. They could then report estimates using models that include or exclude their proposed strategies, and thereby allow the reader to consider her own priors about rival explanations, and for her own inferences about how to update her beliefs.

How should a map be built? In her role as epistemic map maker, a researcher must consider what information readers might need to form inferences. In particular, this implies considering what sets of assumptions should be mapped, and how both assumptions and estimates should be grouped and displayed. Developing a complete protocol for mapping is likely to be an ongoing process, and beyond the scope of this essay. In Appendix A, we synthesize and extend previous proposals to provide some ideas about a native protocol for mapping. Fortunately, in developing this etiquette, we can draw on examples from other disciplines. In psychology and economics, scholars have proposed similar approaches (Simonsohn, Simmons et al., 2015; Leamer, 2010). The best practical examples of protocols for mapping can be found in political science. There, scholars struggling with the issue of model selection have actively taken to "mapping the message". In doing so they have demonstrated how maps can be processed and displayed, and how effective they can be in resolving empirical debates.

Insert Figure 1 about here

A study by three political scientists, Durlauf, Nivarro, and Rivers (2016) (DNR), provides a useful template for epistemic mapping, as well as a demonstration of its potential power to improve the use of testimony in management. In response to an ongoing and angry debate about the effect of laws governing concealed firearrms, DNR mapped the connections between a large number of assumptions and implied estimates laws (Lott and Mustard, 1997; Black and Nagin, 1998; Duggan, 2001. To do this, they first determined the boundary of the set of assumptions to be mapped. After reviewing the relevant literature, they concluded this set should include varying assumptions about the types of crime effected, the timing of change after passage of a new law, the use of conditioning variables, the boundary of the effected population, possible rival hypotheses, and so on (see Figure 1). They then estimated all 864 models spanned by these assumptions, and created multiple maps of the connections between assumptions and the marginal effects implied by each model. Some of the maps they provide summarize the overall robustness of inference across a wide range of specifications;

others allow users to index from preferred assumptions to specific outcomes. These maps fix particular assumptions but allow others to change, and thereby allow readers to observe the robustness of estimates conditional on particular assumptions (see Figure 2). In principle, this process of fixing assumptions could be extended indefinitely. Indeed, an interactive version of DNR's map might allow a reader to index from any particular set of assumptions to the resulting estimates.

The potential for epistemic mapping to allow collective learning and consensus is illustrated by the implications of DNR's analysis. Prior to their article, dozens of scholars had used the same data sources to make opposing claims about the effect of concealed carry laws (Lott and Mustard, 1997; Black and Nagin, 1998; Duggan, 2001). One side opined that "more guns led to less crime", and the other side reached the opposite conclusion. Eventually, the National Research Council decided to intervene, but the committee members themselves could not agree on which assumptions were "correct" (Council, 2005). DNR's analysis revealed why the debate had raged for two decades without conclusion: the inference researchers made depended on their initial assumptions. Had early investigators of the issue published an epistemic map like the one eventually constructed by DNR, readers would have perceived more easily the sensitivity of the results to various researcher choices. Two decades of wasteful debate and misguided public policy might have been avoided. In many areas of management research, where replication is rare or difficult, "mapping the message" may provide the only means for conveying the uncertainty, or at least contingency, of empirical estimates.

Insert Figure 2 about here

IMPLICATIONS

In this review article, we develop a theoretical perspective on learning from testimony about quantitative empirical research in management. We outline the epistemology of testimony and locate vulnerabilities in management reports. We review common types of empirical claims and the conditions required for such claims to be veridical. We assess popular approaches for allowing reduction of testimony but conclude that these approaches are unlikely to be effective in many areas of management research. As a result, we conclude, management scholars must find another way to build shared understanding. We propose that epistemic maps can help researchers and readers avoid the most difficult aspects of testimony's vulnerability problem. In the section below, we discuss implications of our analysis for users of empirical reports, empirical researchers, and developers of theory.

Better use of empirical reports

We hope that our analysis will encourage greater humility about the knowledge we extract from research reports. We, and other scholars, have noted that single "findings" are often accepted as an incontrovertible basis for knowledge (Hubbard, Vetter et al. 1998). It is common, for example, to hear statements such as "[researcher] showed X is associated with Y", or "we know from [researcher], that X is associated with Y". We contend that such statements exaggerate what we really know. As discussed in this article, competence to make such claims requires adherence to a strict set of conditions, and readers can seldom observe whether those conditions have been met. Thus our endorsement of such claims, and our willingness to pass them on, implies either a reasoned choice to adopt a position of presumptive trust, or, more likely, carelessness in language or thought. In either case, we must take care to use language which captures the vulnerability of testimony and our responsibility as stewards of the knowledge we pass on. For example, rather than think (or say), "[researcher] showed that X is associated with Y", we should think "based on the evidence reported by [researcher], I infer support for the hypothesis that X is associated with Y." Taking responsibility for our knowledge in this way will make us better scholars.

Our proposal for epistemic maps makes the reader's responsibility both evident and actionable. By giving the researcher a map of the links between assumptions and outcomes, the reader is encouraged to form her own interpretations. If she has strong priors about the virtues of certain hypotheses or models, she can bring these to the relationships reported in an epistemic map and make stronger inferences. If she has weak priors, she can inspect the full range of outcomes and form weaker interpretations. Either way, the epistemic map makes it clear she must accept that she bears responsibility for the inferences she passes on.

Better research and reporting

A more judicious approach to empirical claims may lead authors to engage in better and more forthright research. At present, authors seem to believe that they must identify interesting patterns of evidence, develop explanations for them, and then test these explanations – all in the same report. The result is that different empirical methods are mixed together in a manner that reduces their effectiveness. Exploratory methods such as abduction implies the consideration of a wide range of alternatives. Frequentist statistical analysis, in contrast, requires just the opposite – strict restraints on the ability of the researcher to explore the data. Combining the two approaches leads to an impoverished research practice that impedes the discovery of good explanations, and yet still fails to justify frequency claims. Better awareness of the conditions for justified claims, and recognition of the value of exploratory research, should allow researchers to choose the style of empirical analysis that best suits their needs and to engage in more forthright reporting of appropriate inferential claims.

Our proposal for epistemic maps could help encourage and guide exploratory research. It is our observation that many empirical studies include some exploratory elements, and yet these elements are often hidden or incompletely reported. Epistemic maps provide a way for reporting such exploration, and encourage researchers to engage in good practices. Creating a map stimulates scholars to think about the major classes of explanations in the set, and how these classes should be displayed. It also endorses the use of what Banerjee et al. (2017) call structured speculation. Freed from the need to marshal a grand defense of a single conclusion, scholars can discuss the merits of different interpretations.

Better Theory Development

Recognition of the difficulty of credibly communicating frequentist and belief claims may assist in the development and refinement of management theory. As Hambrick (2007) discusses, some areas of management fetishize the surfacing and testing of new theory, and journals often ask authors to accomplish both goals in the same paper (Hambrick, 2007). The result is that both erroneous findings and misguided theory proliferate. To allow better grist for theory building, Hambrick (2007) proposes allowing researchers to report what he calls empirical "facts".

Our analysis extends Hambrick's call by considering how such facts are to be interpreted. As we have discussed in this paper, testimony about a "fact", such as an observed association between two variables, is only useful if the author advances claims about which she is competent and sincere. The reader must ascertain the nature of the claim being made and assess whether or not it is veridical.[12] To be useful for theory, facts must be presented in a way that allows their interpretation.

[12] As an example of a useful fact, Hambrick envisions a historical epidemiologist who has a "hunch" that cigarette smoking does "bad things to people" and finds evidence that smoking is "associated with an array of serious maladies". He contends this fact would have helped in the development of theory. We agree, but only if readers had a way to

We agree with Hambrick (p. 1348) that credibly reported facts can allow readers to "direct their efforts at understanding why and how those facts came to be." We would amend his argument by encouraging the reporting of facts in a manner that allows leveraging the private knowledge and creative powers of readers. A lone researcher is limited by her knowledge and insight. A group of readers with access to, and understanding of, a map of contingent "facts" may be better able to imagine fruitful avenues for theory development.

Limitations and Future research

As with any study or review, our research has significant limitations. Our review focuses on the theoretical realm, and thus some of these weaknesses involve the connection between our analysis and practice. Throughout this essay we consider how readers ***should*** respond to testimony, but we provide only limited consideration of how readers currently use reports they read in journals or hear in seminars. It is our observation that a significant number of scholars use default rules of trust or doubt, but other strategies appear to be in use as well. Some scholars trust the research of people with whom they have personal experience. Others try to dig into what backing evidence is provided, in order to determine if there are obvious problems with justification. Future work should explore the prevalence and effectiveness of these heuristics.

We also recognize that our proposed solutions will themselves face conceptual and practical obstacles. For example, our concept of epistemic mapping avoids some problems of testimony, but it leaves others to be contended with. Mapping requires choices about what is included or excluded, as well as scale and detail. While we might hope that these choices will be made following well-established criteria (Tufte, 2001), we must acknowledge that some researchers may use maps persuasively in an attempt to manipulate rather than inform the reader

assess, contemporaneously, its "factness". In hindsight, we know that smoking is harmful, but for years many people thought otherwise. Numerous studies, using different assumptions, were published making a variety of health claims.

(Tyner, 1982). In future research, we hope to develop guidelines for judging the neutrality of epistemic maps.

Future research may also wish to consider how multiple modes of information representation may be combined. Recently, Greve (2018) called for more graphing of data as a critical means for improving testimony. He advocates graphing "the phenomenon to show that there is a meaningful distribution of the outcome one is trying to explain…[and] the distribution of the main independent variables and how they co-vary with the outcome (p. 430)." Such graphing of variables could complement our proposals of epistemic mapping. Greve's graphs provide information on relationships between variables conditional on a set of assumptions. Epistemic maps provide information on how varying assumptions influence estimates of such relationships. In future work, we hope to explore how graphs of different types can be combined.

A final limitation of our analysis is perhaps most significant. To make this essay more tractable, we chose to limit our study to the problem faced by a reader of testimony about *quantitative* empirical research. This means that we consider only one part of the literature. Consideration of the potential for conversational and social mechanisms to allow readers to evaluate the credibility of testimony (c.f. Ketokivi and Mantere, 2010, Mantere and Ketokivi, 2013, Ketokivi, Mantere et al., 2017; Goldman, 2010)) is beyond the scope of our present analysis.

CONCLUSION

We have argued that a reader's ability to reduce frequency and belief claims is impeded, in the management literature, by the difficulty of observing the many assumptions and choices the researcher made in forming such claims. We propose that researchers and readers can avoid this communication problem if researchers report a broad map of connections between empirical assumptions and entailed estimates. Doing so, we contend, would allow readers to use their own assumptions when forming inferences. It is natural to ask how we expect such new practices could come about. Although the question of how to organize a change in the social norms of the research and publication process falls outside the scope of this work, we offer some speculative thoughts here.

A first step, and the main objective of this paper, involves increased awareness about the inherent vulnerability of testimony. We hope that this will make readers or hearers of research more intentional about how they approach empirical reports. This alone, we believe, could influence the practice and reporting of research in management. As predicted by the formal model of Banerjee et al., when readers are undiscerning, researchers will feel less need to ensure that practices and reports are credible. When readers are more demanding, researchers will adjust their practices accordingly. As discussed earlier, readers can also influence the way others interpret research by using language which more accurately expresses the vulnerability of inference from reported research. Well-placed readers, such as reviewers and editors, can go further. They can require that researchers provide epistemic maps that allow advanced claims to be seen in context.

Authors face the more challenging task of adjusting their reporting while still adhering to current norms of publishing. Yet, we believe there are at least two avenues for voluntary action: supplementary epistemic maps and more use of exploratory research. Epistemic maps can be employed regardless of publishing requirements, because modern information technology has made the distribution of supplements virtually free. As van Witteloostuijn notes, “[t]he way we publish is still heavily rooted in practices developed in the 19th century” (Honig, Lampel et al., 2018). Thus, if journals are reluctant to publish such maps, authors can still make them available.

Over time, such voluntary actions with respect to maps may become standardized, in the same way that data disclosure has become a norm at some outlets.

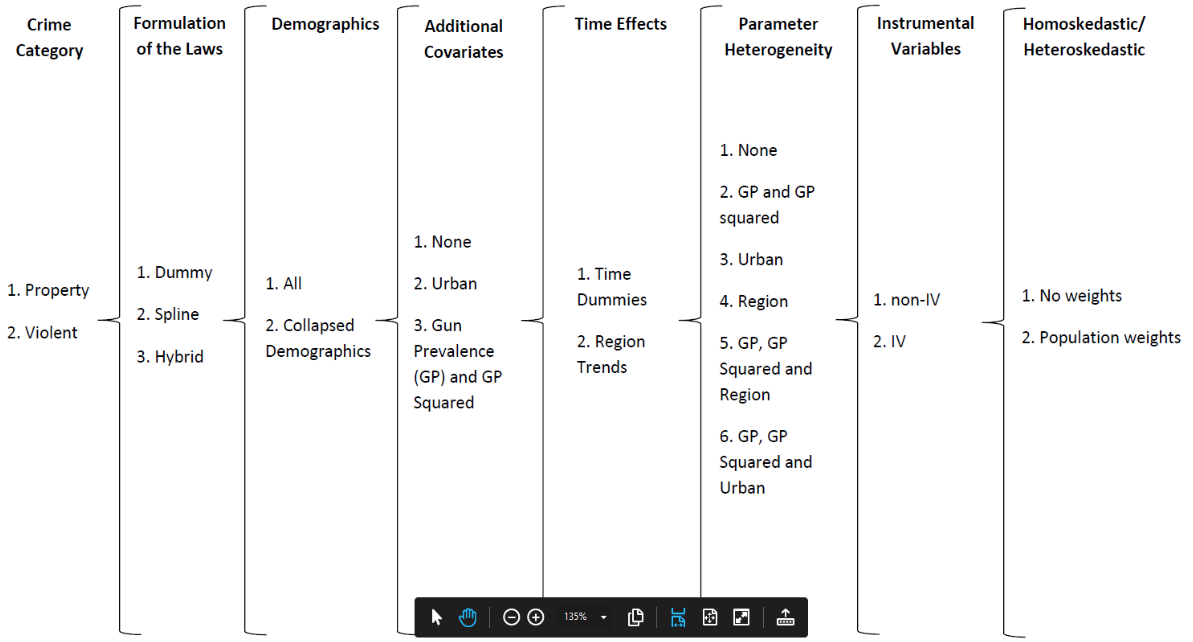
Exploratory search provides another avenue for voluntary action by researchers. Abduction of possible explanations is an integral, indeed, important part of the research and discovery process. It leads to new ideas and theories that might later be subjected to more rigorous testing. And if there is already a great deal of exploration taking place, researchers need only adjust their inferential claims accordingly. There are signs that journals and editors are starting to recognize the value of this approach. For example, The *Academy of Management Discoveries* was created with the idea that we should abduct explanations.

In support of readers and researchers, journal editors could make simple changes in policy. First, they could encourage submission of research that advances preferred explanations backed by epistemic maps. Second, journal editors could educate their reviewers about what they should and should not ask of authors. For example, reviewers should be allowed to ask authors to perform additional tests, map assumptions to findings, and provide data and analytical code for examination. But reviewers should not ask authors to rework a theoretical argument and present it *as if it were specified prior to testing*. Rather, reviewers should ask for ex post description and justification of preferred explanations.

Finally, we hope this article will change perceptions of the problem we face in learning from empirical reports. We do not confront a crisis caused by "bad" researchers who advance "false" reports. Rather we encounter an ongoing and unsolved problem – the vulnerability of testimony as a basis for knowledge. The problem of testimony, like the problem of induction itself, is not solvable in any absolute sense. We contend, however, that we can learn to avoid its most troubling aspects and thereby become better stewards of our collective knowledge.

FIGURE 1

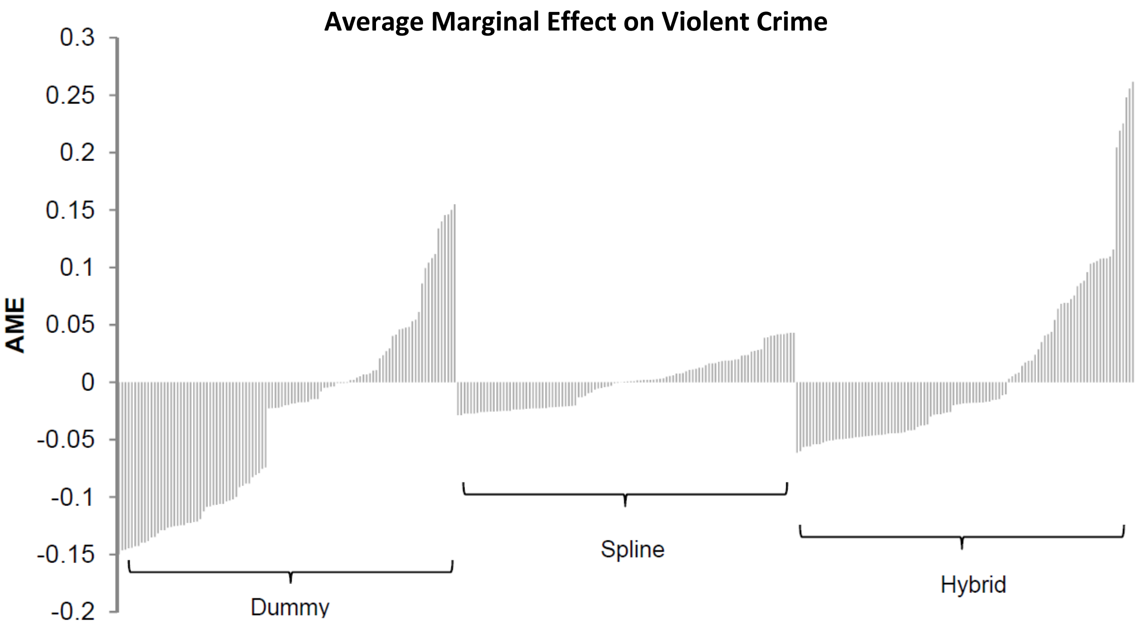
An example of an assumptions space for epistemic mapping. From these, DNR create 864 models for estimating the effect of shall-issue gun laws.



Crime Category	Formulation of the Laws	Demographics	Additional Covariates	Time Effects	Parameter Heterogeneity	Instrumental Variables	Homoskedastic/ Heteroskedastic
1. Property	1. Dummy	1. All	1. None	1. Time Dummies	1. None	1. non-IV	1. No weights
2. Violent	2. Spline	2. Collapsed Demographics	2. Urban	2. Region Trends	2. GP and GP squared	2. IV	2. Population weights
	3. Hybrid		3. Gun Prevalence (GP) and GP Squared		3. Urban		
					4. Region		
					5. GP, GP Squared and Region		
					6. GP, GP Squared and Urban		

FIGURE 2

An example of an epistemic map organized by three assumptions about the functional form of the effect. All other assumptions allowed to vary and specifications sorted, within groups, by effect size.



Appendix A: Guide to Epistemic Mapping

This appendix provides an overview of the process for creating an epistemic map. After providing a simple definition and step-by-step summary, we discuss the selection of “assumption sets” that are the inputs to a map, the selection of summary statistics that are its outputs, presentation of the map itself, and the provision of additional information that could aid interpretation.

In forming our recommendations, we use proposals for related methods. These include “General Systems Analysis” (Leamer 1983), “specification curves” (Simonsohn et al., 2005), “model averaging” (Brock et al., 2007), and “multiverse analysis” (Steegen et al., 2016). All of these proposals start from the observation that empirical researchers make many decisions that are both “arbitrary and defensible” (Simonsohn et al., 2005), and that published research reports rarely show how those decisions influence the outcomes of an analysis. Although the four proposals differ in their details, they all call for increased transparency through a substantial increase in the number of reported estimates. Our approach to epistemic mapping borrows from each of these proposals, and attempts to synthesize them in a way that could be useful for management researchers.

Definition and Process

An epistemic map summarizes the results of multiple analyses that share a common hypothesis.[13] Map users should be able to trace changes in researcher choices and assumptions

[13] Multiple maps might consider alternative hypotheses. For example, DNR map the effect of concealed carry laws on both violent and property crime. See Durloff, Navarro and Rivers (2015), and Brock, Durloff and West (2007) for more technical descriptions.

(e.g. sampling, measurement, specification, identification strategy, etc.) to changes in empirical results. Maps can also provide supplemental information that helps a user identify analyses or assumptions of particular interest, such as the author's preferred model or the analysis that corresponds to assumptions used in a previous study. A major challenge for the epistemic map maker is to choose a set of assumptions broad enough to include any model a reader might wish to evaluate, while preserving the ability to offer a succinct yet comprehensive overview of the relationship between assumptions and results. Before suggesting some guidelines and heuristics for managing this tradeoff, we provide a formal definition of an epistemic map and a simple working example that will help to fix ideas.

An epistemic map is a correspondence between sets of assumptions and summaries of evidence related to a particular claim or hypothesis. Formally, let *A* represent a set of feasible assumptions about an empirical model, and let *a* represent a specific element from that set. Further, suppose that *T* is a set of statistics that summarize a data analysis, with representative element *t*. An epistemic map is a method of summarizing the correspondence *t(a),* which maps each set of assumptions onto an element of T that summarizes the results of the analysis performed under those assumptions.

To make these concepts concrete, suppose we are interested in estimating the relationship between two variables, Y and X, using a linear regression. We can define *A* such that it contains just two assumptions: the functional form of the outcome variable (for simplicity, either logged or in levels) and the inclusion of an additional control variable Z (either yes or no). Furthermore, we can define T as the OLS coefficient estimate on X, along with its standard error. Having made these choices, one representation of the epistemic map can is a simple table, where each row corresponds to a particular set of assumptions and summarizes the results of an analysis based on those assumptions.

Inputs: $a \in A$	*Outputs: $t(a) \in T$*
$Y = \beta X$	$\mathrm{b} = 0.24, \mathrm{se(b)} = 0.13$
$Y = \beta X + \gamma Z$	$\mathrm{b} = 0.22, \mathrm{se(b)} = 0.10$
$lnY = \beta X$	$\mathrm{b} = 0.02, \mathrm{se(b)} = 0.05$
$lnY = \beta X + \gamma Z$	$\mathrm{b} = 0.01, \mathrm{se(b)} = 0.04$

Note: estimates are fictitious.

This type of correspondence provides a way of summarizing model uncertainty – the idea that we do not know which set of assumptions in *A* is actually correct. Although some elements of T, such as the standard error of b, summarize uncertainty based on random sampling or measurement error, this is different from model uncertainty. In particular, the analysis summarized in each row of the table/map takes a particular choice of a ∈ A to be the "correct" assumptions about the functional form of Y and the appropriate set of control variables. In reality, however, authors and readers are typically uncertain about many assumptions, and an epistemic map provides a tool for depicting this model uncertainty when considering testimony about the relationship between Y and X.

Some sets of assumptions are more plausible than others, in the sense that they are more consistent with information that is external to the empirical model, but nevertheless available to the author. The final element of an epistemic map, therefore, consists of auxiliary information that helps a user of testimony assess the relative credibility of different assumption sets. Formally, this information could take the form of probability distribution, p(a), over the elements of A. In practice, we expect it will often take a qualitative form, similar to the current practice in management research of providing *ad hoc* arguments to support the use of a particular set of assumptions as the preferred empirical model for a given study. For instance, in the toy example above, an author might note that they find “log Y” to be a more credible assumption because the variable Y is strictly positive and has a skewed distribution.

Although the simple example shown above would not be especially helpful to readers, it does highlight the four main steps in the mapping process:

1) Selecting Inputs: A
2) Selecting Outputs: T
3) Reporting the Map: t(a)
4) Reporting credibility weights: p(a)

Selecting Inputs to an Epistemic Map

Every decision that goes into an empirical analysis represents a possible dimension of model uncertainty, and every applied empirical researcher knows that there are a large number of decisions that need to be taken. Some common types of decisions include:

- Restrictions on observations to include in the estimation sample
- Choice among alternative measures of outcome and explanatory variables
- Selection of control variables (including fixed effects)
- Selection of functional forms
- Choice of estimation methods and sampling weights
- Choice among alternative identification strategies

Although this is an incomplete list, each bullet implies selecting from a wide range of plausible assumptions. The domain of an epistemic map therefore tends to grow combinatorically as new assumptions are included. In practice, it will typically be impossible for a researcher to map the entire multiverse, and the following considerations are helpful in considering how to define the domain A in a manner that produces a map that is both informative and tractable.

Baseline Etiquette: When considering what functional forms, estimation methods and control variables to include in *A*, there are often some “baseline” values that readers will expect to see reported. For example, mapping etiquette dictates that it is normal to include an OLS regression, and to show how the OLS results change by the addition of the "full" set of control variables. Readers may also have substantive reasons for wanting to see a particular measurement or specification choice. For example, it is more meaningful to measure the outcome of some processes in terms of dollars, as opposed to log-dollars, even if the latter approach leads to a better statistical fit.

Flexibility in Specification Choice: It is sometimes possible to report results from very flexible regression specifications that nest a host of other models, and thereby reduce the magnitude of the combinatorial problem when defining the domain to be mapped. When trying to control for unobserved heterogeneity, it is often a good idea to choose flexible specifications, such as fixed effects, in order to reduce the complexity of the mapping problem. On the other hand, if one is interested in this heterogeneity, it is often preferable to try and measure it directly.[14]

Multiple Measures:_There are often many possible ways to measure a given construct, and maps should generally indicate how results vary with the choice of measures. For example, a time trend may be modeled as year fixed effects, or linear or more flexible polynomial time trend. Each is a mutually exclusive set of controls introduced to model the construct "time". Fortunately, one would not estimate models that include each at the same time and hence introducing new measures, as opposed to new constructs, does not lead to exponential growth in the number of assumption sets.

Theoretically Motivated Controls: One dimension of mapping that can easily lead to an explosion in the size of the domain *A* is the use of alternative sets of control variables, particularly

[14] For example, consider a panel data analysis where a researcher observes many firms, each assigned to a particular industry, over a period of years. One way to control for differences across industries is to include a host of industry-level control variables. An alternative approach is to include industry-by-year fixed effects. The second approach is more flexible because the industry-by-year effects subsume all of the possible combinations of industry-level controls that one might want to explore (technically, the fixed effects are co-linear with any combination of industry-level controls). However, if the researcher is interested in analyzing an industry level construct, use of industry fixed effects would prevent this.

if we include all their interactions. This is also an area where maps could easily be manipulated, for example by including many combinations of uninformative controls, so it appears that hypothesized relationships are very stable. In general, authors should adopt a rule of thumb that any additional control variable, and their interactions, need to be justified with respect to a particular theory, such as an omitted variable or source of unobserved heterogeneity that could potentially bias or confound the parameter(s) of interest. The number of permutations of sets of control variables may nevertheless grow quite large, and while authors could easily report results for every combination in electronic form (e.g. as a database), we expect that in some cases they might reasonably omit some control sets when providing a graphical representation of the map.

Selecting Outputs for an Epistemic Map

There are typically many ways to summarize the evidence obtained from an analysis and choosing a particular set of summary statistics to include is an important part of the epistemic mapping process. For instance, a researcher might report a parameter estimate and its standard error. But she may also report summaries of model fit, such as R-squared or Bayesian Information Criterion (BIC).

In general, what is reported will depend upon the claims being evaluated, which determine the nature of the quantitative analysis performed, and the types of output that it can provide. Many of the principles that apply when summarizing the results of a single quantitative analysis carry over to the problem of deciding what to report. Nevertheless, a map maker should always provide some explanation of how they selected the elements reported, and there are some general guidelines that can inform this choice.

Ease of Comparison: For a map to be easily interpreted, it is important that the summary of each analysis is presented in a manner that facilitates comparison across assumption sets. In that way, a user of the map can easily understand what sets of assumptions provide evidence that support or do not support the claim, and if possible, how strong is the support in each case. For example, researchers should, generally, report marginal effects to allow comparability across models. [15]

Including uncertainty: Most of the quantitative analyses reported in management research provide estimates (e.g. regression coefficients) and also statistics that summarize uncertainty associated with those estimates (e.g. standard errors). This is good practice, and when parameters are part of the map, estimates of parameter uncertainty should also be included. See Figure A.1.

Reporting Maps

Ideally, maps should be easy to read. As Leamer puts it, "[t]he job of a researcher is then to report economically and informatively the mapping from assumptions into inferences." There are many possible ways to summarize the correspondence between assumptions and outputs, however, and a representation that is well suited to one reader or piece of testimony may be poorly suited to another. As with physical cartography, epistemic map makers face a trade-off between summarizing in order to highlight key features and providing more detail to increase fidelity. Because there is no "correct" solution to this problem, epistemic maps can be reported in a variety of ways – and such reporting will generally reflect the researcher's question.

In general, unless prohibited by disclosure constraints, we recommend that authors should publish a database that provides the complete mapping from each set of assumptions, *a*, to the associated summary statistics, t(a). This will allow readers to explore and analyze the map on their

[15] Of course, how marginal effects are calculated needs to be reported as with many functional forms the marginal effects are functions of the data. Depending on the question, it may be advisable to report marginal effects at more than one value beyond the conventional average and / or the average marginal effect.

own. Authors may also choose to report other projections that can aid readers in the understanding of map. For example, authors might want to focus on showing the variability in one element of *t(a)* across the entire domain, *A*. Durlauf, Navarro and Rivers provide similar figures, as illustrated in the paper. Whereas Figure A.1 emphasizes the overall range of estimates, another way to display the information in an epistemic map is to highlight how the range of estimates changes when specific assumptions are turned "on" or "off" using either graphs or tables. Generally, authors will choose to report the results in many ways, as different projections will help clarify the patterns in the map.

At least initially, our expectation is that epistemic maps will be presented in forms that they can be included as part of a traditional journal article. For example, in some fields it is becoming commonplace for empirical papers to contain lengthy appendices detailing a wide range of robustness checks. Epistemic mapping can be viewed as a more thorough and systematic version of ad hoc robustness testing in response to the comments of editors and referees. Building on that idea, maps (and perhaps the associated data) could easily become a standard part of the online appendices liked to a given publication.

Over time, new tools are likely to change the way that maps are reported. For example, researchers might create on-line spaces where readers could plug in assumptions and explore how results change in response. Actual examples of this can be found on political poll-aggregation sites, where readers can exclude or include various polls to see how the aggregate values change. Other researchers are exploring the use of digital "notebooks" that provide readers with access to every input needed to replicate a quantitative analysis – from raw data, through the code used to clean and analyze that data, up to the final summary reports – and create the possibility for interactive epistemic mapping.

Reporting Credibility Weights

Because an epistemic map is a merely a correspondence, it is neutral with respect to the input assumption sets, treating each of them equally. In practice, however, a researcher may see some assumptions as more credible, and wish to give the resulting estimates more weight in the implicit model-averaging process, because those assumptions are consistent with external information. For example, a researcher might consider several ways of measuring a key theoretical construct, but strongly prefer one measure because it has a high degree of inter-rater reliability, or a strong link to the underlying theory, or come from a particularly trusted source.

Information about the credibility of alternative assumption sets can sometimes be included in T. For example, a researcher may wish to report summary measures of model fit, or other statistics drawn from the econometric literature on model selection and specification testing. However, we expect that it will be more common for "credibility weights" to presented separately from the map itself, just as the legend to a physical map is not part of the map *per se*. Much of this presentation may be qualitative, as is the case in many current research papers, where author provide arguments in favor of the set of assumptions behind their baseline empirical model.

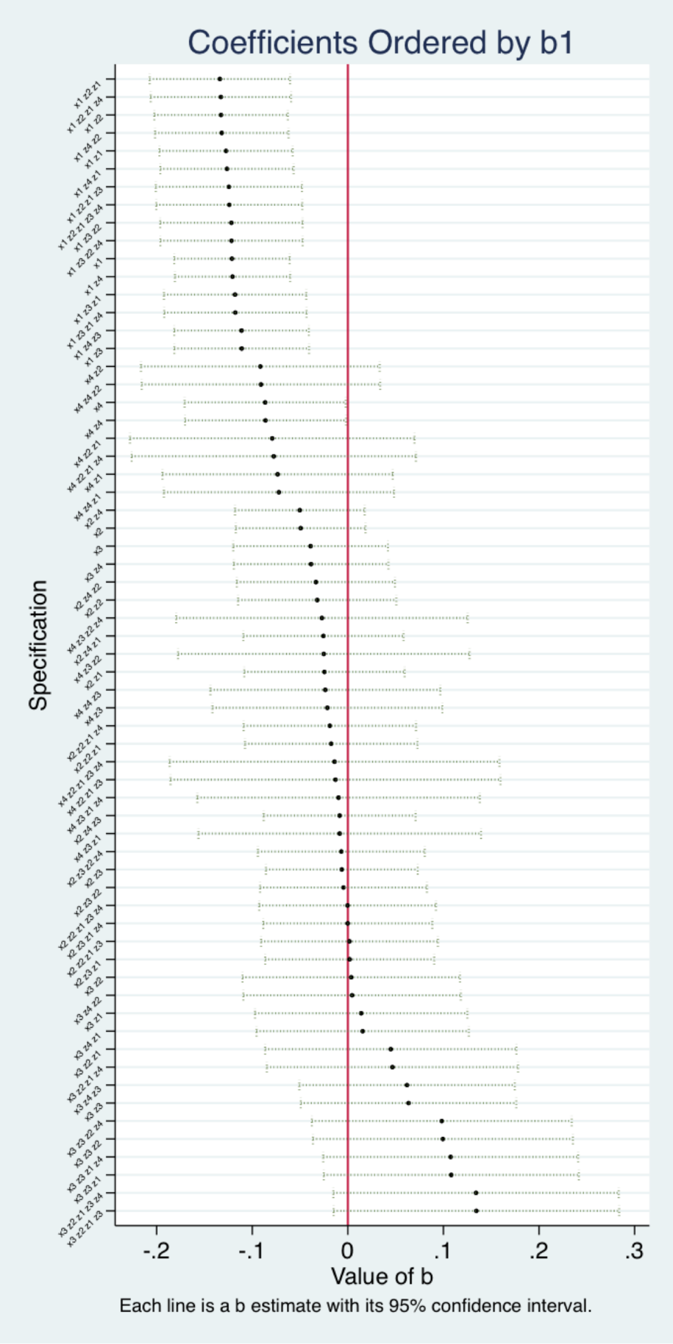
We do not have strong guidelines or preferences related to the representation of credibility weights. In general, we imagine that authors will begin by presenting the entire unweighted map, and then suggest reasons what some parts of the map should be weighted more heavily than others. When reasoning to the best model, authors should include specific reference to the epistemic virtue that justifies their preference.

A special case where this type of supplemental information is particularly important occurs when the analyst wishes to make a belief claim based on a particular identification strategy.

Identification strategies typically make maintained assumptions about the relationship between observed and *unobserved* features of an empirical model. Though it is impossible to provide direct evidence about unobservables, it is typically possible to test implications of the identification assumptions using ancillary data. These tests are important for readers to assess the credibility of the assumptions which are typically necessary conditions for the belief claim to be veridical. Papers seeing to make causal inferences or advance belief claims, therefore, should always provide some sort of legend that helps readers assess maintained assumptions behind the identification strategy that is used to rule out alternatives to a preferred hypothesis.

FIGURE A.1:

An Example of How an Epistemic Map Could Be Reported[16]



[16] The specifications are reported on the *Y-axis,* and ordered by the size of the measured coefficient on x_i and bounded by their 95% confidence intervals. The zero line is highlighted vertically in the figure.

REFERENCES

Adler, J. (2006). Epistemological problems of testimony. *The Stanford Encyclopedia of Philosophy.* . Stanford, CA, The Metaphysics Research Lab, Stanford University.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: an empiricists guide*. Princeton, NJ, Princeton University Press.

Angrist, J. D. and J.-S. Pischke (2010). "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of Economic Perspectives* **24**(2): 3-30.

Antonakis, J., S. Bendahan, P. Jacquart and R. Lalive (2010). "On making causal claims: A review and recommendations." *The leadership quarterly* **21**(6): 1086-1120.

Athey, S. and G. Imbens (2015). " A Measure of Robustness to Misspecification." *The American Economic Review,* **105**(5): 476-480.

Banerjee, A. V., S. Chassang and E. Snowberg (2017). Decision Theoretic Approaches to Experiment Design and External Validity. *Handbook of Economic Field Experiments*, Elsevier. **1:** 141-174.

Bartlett, T. (2017). "Spoiled science." *Chronicle of Higher Education* **63**(28).

Bateman, B. W. and Philosophy (1987). "Keynes's changing conception of probability." *J Economics* **3**(1): 97-119.

Behfar, K. and G. A. Okhuysen (2018). "Discovery Within Validation Logic: Deliberately Surfacing, Complementing, and Substituting Abductive Reasoning in Hypothetico-Deductive Inquiry." *Organization Science* **29**(2): 323-340.

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

Bettis, R. A., S. Ethiraj, A. Gambardella, C. Helfat and W. Mitchell (2016). "Creating repeatable cumulative knowledge in strategic management: A call for a broad and deep conversation among authors, referees, and editors." *Strategic Management Journal* **37**(2): 257-261.

Black, D. A. and D. S. Nagin (1998). "Do right-to-carry laws deter violent crime?" *The Journal of Legal Studies* **27**(1): 209-219.

Carnap, R. (1962). *Logical Foundations of Probability*. Chicago, IL, The University of Chicago Press; 2nd edition (1962).

Chatterjee, A. and D. C. Hambrick (2007). "It's all about me: Narcissistic chief executive officers and their effects on company strategy and performance." *Administrative Science Quarterly* **52**(3): 351-386.

Coady, C. A. (1992). *Testimony: A philosophical study*. Oxford, UK, Clarendon Press.

Council, N. R. (2005). *Firearms and violence: a critical review*, National Academies Press.

Davis, M. S. (1971). "That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology." *Philosophy of the Social Sciences* **1**(2): 309-344.

De Long, J. B. and K. Lang (1992). "Are all economic hypotheses false?" *Journal of Political Economy* **100**(6): 1257-1272.

Douglas, H. (2000). "Inductive risk and values in science." *Philosophy of Science* **67**(4): 559-579.

Douven, I. (2002). "Testing inference to the best explanation." *Synthese* **130**(3): 355-377.

Duggan, M. (2001). "More guns, more crime." *Journal of Political Economy* **109**(5): 1086-1114.

Durlauf, S. N., S. Navarro and D. A. Rivers (2016). "Model uncertainty and the effect of shall-issue right-to-carry laws on crime." *European Economic Review* **81**: 32-67.

Edwards, A. W. F. (1984). *Likelihood*, Cambridge, UK, Cambridge University Press.

Elgin, C. (2002). "Take If from Me: The Epistemological Status of Testimony." *Philosophy and Phenomenological Research* **65**(2): 291-308.

Everett, J. A. C. and B. D. Earp (2015). "A tragedy of the (academic) commons: interpreting the replication crisis in psychology as a social dilemma for early-career researchers." *Frontiers in Psychology* **6**: 1152.

Fisher, R. (1960). "The design of experiments ((1935, 1st)." *Edinburgh: Oliver and Boyde*.

Fricker, E. (1994). Against gullibility. *Knowing from words*. New York, NY, Springer**:** 125-161.

Fricker, E. (1995). "Critical notice." *Mind* **104**(414): 393-411.

Fricker, E. (2002). "Trusting others in the sciences: a priori or empirical warrant?" *Studies in History Philosophy of Science Part A* **33**(2): 373-383.

Fricker, E. (2004). Testimony: Knowing through being told. *Handbook of Epistemology*, New York, NY, Springer**:** 109-130.

Gelman, A. (2014). "Preregistration: what's in it for you?"   Retrieved August, 28, 2019, from https://statmodeling.stat.columbia.edu/2014/03/10/preregistration-whats/.

Gelman, A. (2015). "Statistics and the crisis of scientific replication." *Significance* **12**(3): 23-25.

Gelman, A. and G. Imbens (2013). Why ask why? Forward causal inference and reverse causal questions, National Bureau of Economic Research.

Gelman, A. and E. Loken (2016). "The statistical crisis in science." *J The best Writing on Mathematics* **2015**: 305.

Glymour, C. (1980). Why I am not a Bayesian. *Theory and Evidence*. Princeton, NJ, Princeton University Press. .

Goldfarb, B., & King, A. A. (2016). Scientific apophenia in strategic management research: Significance tests & mistaken inference. *Strategic management journal, 37*(1), 167-176.

Goldman, A. (2010). "Systems-oriented social epistemology." *Oxford Studies in Epistemology* **3**: 189-214.

Greve, H. R. (2018). "Show Me the Data! Improving Evidence Presentation for Publication." *Management Organization Review* **14**(2): 423-432.

Hacking, I. (1965). *The Logic of Statistical Inference*. Cambridge, UK, Cambridge University Press.

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic Desk Examination Edition*. Cambridge UK, Cambridge University Press.

Hambrick, D. C. (2007). "The field of management's devotion to theory: Too much of a good thing?" *Academy of Management Journal* **50**(6): 1346-1352.

Hamilton, B. H. and J. A. Nickerson (2003). "Correcting for endogeneity in strategic management research." *Strategic Organization* **1**(1): 51-78.

Hardwig, J. (1985). "Epistemic dependence." *The Journal of Philosophy* **82**(7): 335-349.

Heckman, J. J. (2008). "Econometric causality." *International Statistical Review* **76**(1): 1-27.

Heckman, J. J. and B. Singer (2017). "Abducting Economics." *American Economic Review* **107**(5): 298-302.

Honig, B., J. Lampel, J. A. Baum, M. A. Glynn, R. Jing, M. Lounsbury, E. Schuessler, D. G. Sirmon, A. S. Tsui and J. P. Walsh (2018). "Reflections on Scientific Misconduct in Management: Unfortunate Incidents or a Normative Crisis?" *Academy of Management Perspectives* **32**(4): 412-442.

Hubbard, R., D. E. Vetter and E. L. Little (1998). "Replication in strategic management: Scientific testing for validity, generalizability, and usefulness." *Strategic Management Jour*nal 19(3): 243-254.

Hume, D. (1748/1993). *An enquiry concerning human understanding*. Indianapolis/Cambridge, Hacket Publishing Co.

Ioannidis, J. P. (2005). "Why most published research findings are false." *PLoS medicine* **2**(8): e124.

John, L. K., G. Loewenstein and D. Prelec (2012). "Measuring the prevalence of questionable research practices with incentives for truth telling." *Psychological Science* **23**(5): 524-532.

Ketokivi, M. and S. Mantere (2010). "Two strategies for inductive reasoning in organizational research." *Academy of management review* **35**(2): 315-333.

Ketokivi, M., S. Mantere and J. Cornelissen (2017). "Reasoning by Analogy and the Progress of Theory." *Academy of Management Review* **42**(4): 637-658.

Kidwell, M. C., L. B. Lazarević, E. Baranski, T. E. Hardwicke, S. Piechowski, L.-S. Falkenberg, C. Kennett, A. Slowik, C. Sonnleitner and C. Hess-Holden (2016). "Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency." *PLoS biology* **14**(5): e1002456.

Leamer, E. E. (1983). "Let's take the con out of econometrics." *The American Economic Review* **73**(1): 31-43.

Leamer, E. E. (2008). *Macroeconomic patterns and stories*. Berlin, GER, Springer Science & Business Media.

Leamer, E. E. (2010). "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives* **24**(2): 31-46.

Lipton, P. (2003). *Inference to the best explanation*. Abingdon, United Kingdom, Routledge.

Longino, H. (2002). The social dimensions of scientific knowledge. *The Stanford Encyclopedia of Philosophy*. Stanford, CA, The Metaphysics Research Lab, Stanford University.

Lott, J., John R and D. B. Mustard (1997). "Crime, deterrence, and right-to-carry concealed handguns." *The Journal of Legal Studies* **26**(1): 1-68.

Lounsbury, M. and C. M. Beckman (2015). "Celebrating organization theory." *Journal of Management Studies* **52**(2): 288-308.

Mackonis, A. (2013). "Inference to the best explanation, coherence and other explanatory virtues." *Synthese* **190**(6): 975-995.

Manski, C. F. (1995). *Identification problems in the social sciences*, Cambridge, MA, Harvard University Press.

Mantere, S. and M. Ketokivi (2013). "Reasoning in organization science." *Academy of Management Review* **38**(1): 70-89.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*, Chicago, IL, University of Chicago Press.

Mayo, D. G. and D. R. Cox (2006). Frequentist statistics as a theory of inductive inference. *Optimality*, Institute of Mathematical Statistics**:** 77-97.

Mayo, D. G. and A. Spanos (2011). Error statistics. *Philosophy of statistics*, Elsevier**:** 153-198.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1), 235-245.

Li, F. and F. Mealli (2014). "A Conversation with Donald B. Rubin." *Statistical Science* 29(3): 439-457.

Psillos, S. (2002). Simply the best: A case for abduction. *Computational logic: Logic programming and beyond*, New York, NY, Springer**:** 605-625.

Rosenthal, R. (1979). "The file drawer problem and tolerance for null results." *Psychological Bulletin* **86**(3): 638.

Sanborn, A. N. and T. T. Hills (2014). "The frequentist implications of optional stopping on Bayesian hypothesis tests." *Psychonomic Bulletin & Review* **21**(2): 283-300.

Schmidt, S. (2009). "Shall we really do it again? The powerful concept of replication is neglected in the social sciences." *Review of General Psychology* **13**(2): 90-100.

Schneider, J. W. (2015). "Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations." *Scientometrics* **102**(1): 411-432.

Schurz, G. (2008). "Patterns of abduction." *Synthese* **164**(2): 201-234.

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

Senn, S. (2011). "You may believe you are a Bayesian but you are probably wrong." *Rationality, Markets Morals* **2**(48-66): 27.

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

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

Simonsohn, U., L. D. Nelson and J. P. Simmons (2014). "P-curve: a key to the file-drawer." *Journal of Experimental Psychology: General* **143**(2): 534.

Simonsohn, U., J. P. Simmons and L. D. Nelson (2015). "Specification curve: Descriptive and inferential statistics on all reasonable specifications." *Available at SSRN 2694998*.

Sine, W. D. and B. H. Lee (2009). "Tilting at windmills? The environmental movement and the emergence of the US wind energy sector." *Administrative Science Quarterly* **54**(1): 123-155.

Spanos, A. (2010). "Is frequentist testing vulnerable to the base-rate fallacy?" *Philosophy of Science* **77**(4): 565-583.

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

Stove, D. C. (2014). *Popper and after: Four modern irrationalists*, New York, NY, Pergamon Press.

Tatsioni, A., N. G. Bonitsis and J. P. Ioannidis (2007). "Persistence of contradicted claims in the literature." *Journal of the American Medical Association* **298**(21): 2517-2526.

Tihani, Lazlo., Learning and Reporting after the Replication Crisis, Academy of Management Annual Conference, Boston, August 12.

Tufte, E. R. (2001). *The visual display of quantitative information*, Graphics press Cheshire, CT.

Tyner, J. A. (1982). "Persuasive cartography." *Journal of Geography* **81**(4): 140-144.

Van Riel, R. and R. Van Gulick (2016). "Scientific Reduction." *Stanford Encyclopedia of Philosophy*.

Wilholt, T. (2013). "Epistemic trust in science." *The British Journal for the Philosophy of Science* **64**(2): 233-253.

Young, N., J. Ioannidis and O. Al-Ubaydli (2008). "Why current publication practices may distort science." *PLoS medicine* **5**(10): e201.

Zaheer, A. and G. Soda (2009). "Network evolution: The origins of structural holes." *Administrative Science Quarterly* **54**(1): 1-31.