Causal Inference in Strategy Research: Lovely or Likely?

Comments for Utah 2024 Strategy Summit
Panel on Causal Inference in Strategy

Brent Goldfarb, University of Maryland
David A. Kirsch, University of Maryland
Sandeep Pillai, Bocconi University

July 29, 2024

> A good business strategy deals with the edge between the known and the unknown. Again, it is competition with others that pushes us to edges of knowledge. Only there are found the opportunities to keep ahead of rivals…. Given that we are working on the edge, asking for a strategy that is guaranteed to work is like asking a scientist for a hypothesis that is guaranteed to be true—it is a dumb request. The problem of coming up with a good strategy has the same logical structure as the problem of coming up with a good scientific hypothesis. The key differences are that most scientific knowledge is broadly shared, whereas you are working with accumulated wisdom about your business and your industry that is unlike anyone else's. A good strategy is, in the end, a hypothesis about what will work. Not a wild theory, but an educated judgment. And there isn't anyone more educated about your businesses than the group in this room.
>
> Richard Rumelt, *Good Strategy Bad Strategy: The Difference and Why It Matters* (p. 242). Crown. Kindle Edition.

Problem

How far can we push toward causal identification in a field focused on novelty, uniqueness, and complementarity? A growing chorus of authors, and indeed the premise of this panel, argue that core questions in Strategy cannot be answered using traditional causal identification techniques. To consider this question, we first note that novelty, uniqueness, and complementarity are not the only challenges we face in applying causal identification techniques to Strategy. Rather, before we consider these Strategy-specific questions, we need to also ask if our applied social science toolkit can deliver the types of answers we seek. We consider these challenges carefully, and then reflect on

if and how new thinking in epistemology might be applied so as to lead to strategy research that better informs strategic design.

As a desk clearing exercise, it is important to be clear as to what we mean by causal inference. We define causal inference in terms of invertibility - a mechanism is causally identified if there is a 1-1 correspondence between the proposed mechanism and the data (Lewbel 2019). In words: the data observed could have been realized if and only if causal mechanism M_i is operating.[1] If and only if is key: identification requires ruling out all other mechanisms. Obviously, we need criteria by which to decide when a claim that M_i is operating to the exclusion of all other mechanisms $M_{j\neq i}$ is warranted. That is, when we should believe there is a 1-1 correspondence. Ronald Fisher's concept of identifying a causal effect hinges on the ability to repeat an experimental manipulation and reliably observe the same outcome (Mayo 2018). Fisher's approach is helpful to the extent that scholars in Strategy seek reliable tools to not only to analyze firms' strategic choices, but also to makereliable recommendations.

The use of causal identification techniques requires relying on three interrelated assumptions - a three legged stool of sorts. Unhappily, two of the legs are flat out broken, and the third is held together by some masking tape and glue. The first stool leg requires that econometric methods identify mechanisms that inform strategic decisions. This is the question that opens this essay. The second leg requires a reliable criteria by which to decide whether a claim of identification is warranted, conditional on a particular assumed data generating process. The third leg brings focus to the problem that causal identification techniques are very challenging to implement without some data peeking, but reliability of the claim-warrant requires proper modeling of the data generating process (DGP), and it challenges criteria we use that require knowing the DGP before collecting the data.

The first leg of the stool is the assumption that average effects are useful in informing strategic decisions. The alternative is that "novelty, uniqueness, and complementarity" implies that the field of strategy is, or should be, the study of questions that are not directly answerable using statistical methods. Rumelt's (2011) reference to the "edges of knowledge" reflects a claim that the context of decisions are particularly important in strategy. This definition of what makes a decision strategic is consistent with, indeed the title of Leiblein, Reuer, and Zenger (2018)'s definition of strategic decisions as those whose outcomes are characterized by interdependence with other

[1] Econometrically, we develop techniques to estimate y=bx+e that justify a claim that if we were to change x this would **cause** a change in y. This cause is a mechanism

contemporaneous firm decisions, competitor decisions, and future decisions. In practice, this implies that decisions that are strategic are unique and reflect the particular context in which they are made. Pillai et al. (2020, 2024) argue that only deep contextual or historical analysis will allow meaningful causal claims in Strategy, but such events – and the decisions taken in the face of them – are, again, difficult to generalize from. Causal identification requires samples of actors facing similar interdependencies to estimate average effects, and out of sample predictions requires stability in interdependencies across space and time. If so, then Local Average Treatment Effects (L.A.T.E.) will be of interest. If not, L.A.T.E. estimates may have little relevance outside these samples. As Leiblein et al. write (p. 562): "[i]n a world where all strategic decisions depend on many other choices, average effects offer rather little value to the decision maker. Therefore, the existence of these decision interdependencies, including higher order interactions and feedback loops, calls regression approaches into question (Bettis and Blettner 2020)." Rumelt said it in so many words. Except in rare cases, traditional regression analysis – that principally allows us to extrapolate from a sample to a population – is fundamentally about unpacking past strategic decisions. For the prospective strategist, Leg 1 is broken.[2]

But even if we stipulate that the study of (past) average effects of strategic choices is of great interest, can we estimate such effects reliably? Will such estimates generalize beyond the sample of study? . Current methods make reliable, general claims exceptionally challenging. Thus, the second leg of the stool, the "prespecification leg", breaks when we recognize that Frequentist statistics are not reliable unless we have strict prespecification. However, coming up with reasonable models will generally require interaction with the data. When interaction with the data is necessary to come up with plausible identification strategies, the methods will not produce reliable statistics (A. King, Goldfarb, and Simcoe 2021; S. D. Pillai, Goldfarb, and Kirsch 2024; Leamer 1983; Bettis 2012; Goldfarb and King 2016). Thus, our goal of measuring the likelihood that observed results are due to chance is put out of reach.

How strict must pre-specification be? A quick glance at FDA randomized control trial drug approval protocols clarifies that ensuring reliability requires a level of precision in pre-specification rarely feasible in Strategy: Researchers must have no flexibility before data collection begins, ensuring complete transparency and predetermined methodologies, including stating hypotheses,

[2] There are many authors I could have leaned on to support this reasoning. Guzman (2022) makes the same case based on a different set of seminal thinkers in strategy (and engages in a heroic attempt to rescue traditional methods given this problem). Of course, one obvious solution is to focus on different questions under the light of the lamp post. However, we are uninterested in such an approach.

defining measures, outlining sampling plans and stopping rules, and specifying analysis techniques. These requirements ensure that the data follow the theory, and allows frequentist statistics to be interpreted at face value. Failure to adhere to these rigorous standards compromises the reliability of research findings and requires the (almost always heroic) assumption that no part of the process was informed by patterns seen in the data.

Many scholars continue to bridle under this stricture; surely, some say, looking at data does not change the calculation of a test statistic. Maybe so, but peeking does change which test statistics we choose to calculate. In particular, it makes us more likely to calculate and report relationships that are precisely estimated in the sample (Goldfarb and King 2016). The degree in which this problem affects the statistics we calculate is idiosyncratic to sample and researcher and is a function of the host of coding, cleaning, and collection judgment calls that are necessary to process our data sufficiently to make sense of it (and this is prior to careful consideration of identifying assumptions as when we consider the third leg momentarily). However, this intractability robs the researcher of the ability to interpret what a given test statistic *means* when we are seeking to establish causal relationships (Mayo 2018, 200–201). Therefore, Leg 2 is also broken.[3]

The third leg, the DGP leg, is that interpreting the second leg's statistical relationships as causal requires identifying assumptions (e.g., exclusion restrictions when employing instrumental variables, parallel trends for a difference in differences estimator, etc.). Methods to map assumptions to consistent estimates are the focus of econometric contributions to the causal inference literature. When researchers question these assumptions, we get an "identification party".[4] Is the exclusion restriction plausible? Are the back doors closed? Lee and Ryall (2024) argue that Strategy scholars sometimes are insufficiently precise in their description of these assumptions, and this sets the stage for particularly vibrant and contentious identification parties. As a remedy, they recommend the use of Directed Acyclical Graphs (D.A.G.s) to force more clarity around the author's assumed data generating process (i.e., model). The hope is that while we can never know if a D.A.G. is the true model, nor can we know if identifying assumptions obtain, we will be more likely to model DGPs correctly, and also more informed conversations can be had if we can precisely articulate our assumptions and then condition our conclusions on these assumptions.

[3] How strict is strict? The point is that the data analysis should be severely constrained prior to collecting the data, and that there be pre-specified data collection and data analysis stopping rules. If we can approach this, it may be the case that statistics in such cases are more reliable than without these constraints. Certainly this should be an input into the criteria for Severe Testing below. Judgments abound.

[4] One of us seems to recall Joanne Oxley attributing this phrase once to Anita McGahan.

No doubt this leg of the stool can be reinforced by D.A.G.s and other forms of data-generating-process TLC. But the hatchet is sharp and subtle. D.A.G.s are difficult to imagine without undermining our ability to evaluate statistical fit. Pillai et al. (2024) highlight the co-dependence between this DGP-leg and the pre-specification leg. Insight into the most likely DGP will come from interaction with the data, but that interaction will compromise the reliability of test statistics as researchers select tests based on statistical fit. The hatchet blade subtly slices the third leg because the researcher's ability to find a plausible causal mechanism compromises their ability to evaluate if this mechanism fits the data.

So where are we? It seems all doom and gloom! Strategy is fundamentally interested in identifying causal mechanisms, and our three-legged causal inference stool lacks legs. But perhaps greater modesty will allow us to still make progress. We propose that a more limited causality may be inferred, and this may yet be useful for questions of interest to strategists. Bits and pieces of mechanisms may be put together and used to shore up the reliability of predictions of the outcomes of strategic decisions. Perhaps, this is what the field is already doing. We argue that facing these flaws in both our epistemology and corresponding testimony head on can lead to a better path of useful knowledge production in Strategy.

Our proposed path forward, admittedly in its first steps, builds on principles and prior work. We rely on our prior work on testimony and abduction (A. King, Goldfarb, and Simcoe 2021, S. D. Pillai, Goldfarb, and Kirsch 2024, Sandeep Devanatha Pillai et al. 2024) and strengthen it with ideas from an epistemology called "Severe Testing" (Mayo, 2018), described below. Rebuilding the broken stool requires a patchwork appropriate for the multidisciplinary field of Strategy. In a sentence: we proscribe much greater modesty in our claims, combined with an aspiration for "severity" in our testing of causal links. We may then use these links as building blocks, accounting for the severity in which they have been tested. Thus, we will repurpose links as "micro-links" to increase the reliability of strategies we design.

Solution: Building a new Stool

To move forward we adopt Mayo & Cox's (2006) idea of "severe testing". In many ways Mayo and Cox integrate what we already do in practice, and helpfully clean up the epistemology. A hypothesis is severely tested when "[d]ata x_0 in test T provide good evidence for inferring H (just) to the extent that H passes severely with x_0, i.e., to the extent that H would (very probably) not have survived the test so well were H false." This formulation sits soundly in Popper's falsification camp,

but asks for an additional step of saying that if we subject *H* to tests that it should fail if it is false, and it passes those tests, we might as well act as if *H* is true.

Severe testing provides an epistemic rule by which to accept causal inferences. Surviving a severe test implies comparing to other possible *H*'s (Mayo 2018) - or more precisely, evaluating the degree in which each *H* is consistent with the sample. Severe testing means that after such an evaluation, we would accept *H* if and only if we would be very likely to observe the patterns in the data if *H* were true, but also if the patterns in the data would be very unlikely if any other *H* were true. In well-executed causal identification strategies that meet Lewbel's invertibility, the competing *H*'s are confined to random chance and quantified by test statistics, such as a p-value. It's a beautiful solution … or it would be if Stool Legs 2 and 3 were intact and we could reliably interpret test statistics![5]

Mayo and Cox deliberately left wiggle room through their choice of words "severely" and "very probably". This elevation of the role of researcher judgment, as Mayo (2018) documents extensively, aligns severe testing with both Fisher's as well as Neyman's and Pearson's insistence that interpretation of statistics ought rest on judgment as opposed to hard rules.[6]

We are about to get a little heterodox, in that Mayo (2018, 200–201) is skeptical that anything can be deemed "very probable" with unreliable statistics. But with our applied researcher's lens, we disagree. In practice, our current practice often includes many characteristics of severe testing, including many that are not quantifiable but "very probable". This includes necessary theoretical assumptions that map descriptions of context to model parameters, theoretical frameworks that are assumed to have predictive value, the weighting of this or that robustness test, and newer attempts at Leg 3 repair in the form of partial identification (see Frake et al. 2023 for a review and application in

[5] The problem of Broken Legs 2 and 3 has led leading thinkers to argue that causal identification is only possible with Randomized Controlled trials. As King et al. (2021: p. 472) summarize: "Angrist and Pischke (2010: 4) argued that "design-based studies," which typically rely on natural experiments, "are distinguished by their prima facie credibility." Other scholars have gone further. For example, statistician Donald Rubin argued that only randomized control trials allow causal belief claims. He described claims based on other approaches, such as tests of possible causes of an observed effect, 'more of a cocktail conversation topic than a scientific inquiry' (as cited in Li & Mealli, 2014: 446)."

[6] Mayo insists that we can use test frequentist test statistics to quantify the severity. She provides a formal accounting of this, but we can suffice with the intuition that we can learn from the precision of estimates, and then quantify how likely a particular truth with p-values, and thus use this to rule out values, as opposed to simply saying that we fail to reject the null. We agree insofar as we are able to pre-specify (Stool Leg 2 remains intact).

Strategy)[7], or epistemic mapping (A. A. King 2023).[8] However, Broken Leg 2 (pre-specification) implies that the assessment of "very probably" cannot rely on statistics alone.[9]

Though we cannot interpret statistics at face value, we can still use them to assess fit in a sample. Thus, our disagreement with Mayo is really a statement that we can make more modest claims. Mayo characterizes severity as severe and weak. Severe requires reliable and interpretable tests. Weak severity is, in Mayo's view, no test at all.[10] If we look no further than our statistics, and prior theory, we may agree with Mayo. But we can push further than this dichotomy. We can use additional information to come to judgments on the degree of severity in our inference - or alternatively, how *likely* it is that a theory explains the data. With this, we are in the world of abduction, or inference to the best explanation - but with a more precise concept for judging *likely*. Hence, we leverage Pillai et al. (2024a, p. 4)'s distinction between explanation and theory: "An explanation refers to a retrospective account of a specific set of observations in a particular context. Explanation is related to, but distinct from theory, which refers to a more general set of principles that abstracts away from a context with the aim of explaining a wide range of phenomena or similar sets of observations across different settings ... An explanation, in contrast, is a retrospective account of a specific set of observations. An explanation is more concrete or practical, whereas a theory might abstract away from a context."

Explaining (past) data as opposed to testing theory allows us to maintain much of our current research process, and repurposes the broken shards of Legs 2 and 3 to fashion a different and more modest stool. It also requires adjusting our testimony to more modest claims (Sandeep Devanatha Pillai et al. 2024). (We will keep the stool from falling with the remnants of the first leg in a bit). Because given that we need to interact with the data to understand it, we cannot completely

---

[7] Partial Identification asks what happens if identifying assumptions are somewhat wrong (see Frake et al. 2023 for a review and application in Strategy). Thus, under some conditions, one can ask what happens if the identifying assumption is a bit wrong.

[8] Epistemic mapping (A. King, Goldfarb, and Simcoe 2021) leans into the model uncertainty by asking if a similar association can be found under differing modeling assumptions (See Andrew King's work on testimony and co-authors, including myself). This is a powerful approach to the problem, though it may be somewhat rare to get the same answer to very broad modeling assumptions. An interesting path forward may be the combination of partial identification and mapping techniques, though that is beyond the scope of this essay. It may be possible, in some circumstances, to repair the third leg sufficiently by generating epistemic maps that demonstrate a relationship insensitive to all plausible models.

[9] There may be lovely exceptions due to some asymmetry. If our estimates are overly precise, we should expect that generally, we are overly likely to reject the null. If we find it very difficult to reject the null, we still may conclude that there is no evidence of a phenomenon in our sample, and if we can convince ourselves that we have a reasonable DGP and our measures are not too noisy, then we might generalize to the population.

[10] Mayo labels this B-E-N-T (Bad Evidence No Test). A failure to rule out alternative hypotheses is not severe testing and hence BENT. Mayo goes further and implies that confirmation based approaches that look no further than showing evidence consistent with a hypothesis is BENT.

refrain from doing so. This compromises our statistics, but we can still explain where there are correlations in our data (and still do our causal econometric tricks). That is, while we can no longer say how likely the patterns in the data are given our theory, we can answer the question of whether the data are even plausible given an explanation. This leverages an asymmetry in the problem of peeking at the data. It should make it easier to find patterns. A failure to find a pattern predicted by an explanation even after searching may be decent evidence the explanation poorly explains the data. Success implies that the explanation is a possibility, but not much more than that.

But settling with the statement that an explanation is a possibility is unsatisfying. And we do not hold that an honest recognition of the tradeoff between getting the right model and our ability to test it within a dataset is a license for an unconstrained exercise in data mining. Our goal remains to test an explanation as severely as possible so as to probe to what degree we might claim thatt a pattern in the data is very unlikely to be coincidental, and that the mechanism causing this pattern is very likely to be M_i. However, with an inability to create an accurate test, we need to triangulate in with multiple sources of information.

First, our search should be guided by an application of all (not just preferred) plausible explanations. Without invertibility, we need to then consider more nuanced evidence and comparisons across theories. A variety of tools and evidence to justify claims as to what best explains these correlations. Explanations have predictions, and we should aspire to look for signs of these predictions in the theory. These predictions should be fair tests of an explanation - and by fair we mean that we should aspire to find tests that reflect the principle of severity: the predictions should *very probably* **not** bear out in the data if the explanation is false. And others should very probably bear out if and *only* if the explanation is true. Severe tests should be severe.

This perhaps gives more clarity in judgments of *likely* in the context of abduction, where explanations are generated, and one is determined "best". Modern abduction is also labeled "inference to the best explanation" (Douven 2022). But "best" requires careful consideration of likely. And although abduction lacks a universal criteria for adjudicating which explanation is most likely, Cox and Mayo's Severe Testing provides one as we apply abduction to our field. Thus, we can use severe testing in a way that acknowledges the process of abduction, recognizes the iterative process of explanatory calibration where we begin with an initial explanation, look for a corresponding pattern, fail to find it, think harder that maybe something other causal mechanism better explains the data, dream up a consequent of this better explanation, check if it is true, and so on.

But if we cannot conduct reliable, severe statistical tests, we then need other criteria to implement the severe testing tool. Pillai et al. (2024a) make the case that historical methods provide a key set of principles by which to do this. In particular, there we argue that we should rely on three principles when evaluating evidence: First, why we observe the evidence in the first place (source criticism), what were the perspectives and beliefs of the decision-makers (hermeneutics), and what was the context of these decisions (contextualization). A severely tested explanation will be highly *consilient* with a range of observations - quantitative, archival, statistical, documentary, testimonial etc. - that emerge from the historical record (we use the term historical liberally as all archival studies are studies of the past) - but our judgment of this consilience requires considerations of the source of the observations, the specific context of the decisions, and consideration of how the decision makers thought and what they were trying to achieve. *Consilience* is an epistemic virtue - a catchall for "fitting to many facts". Severe testing requires developing tests to evaluate whether observations are likely given an explanation, and tests that ignore motivations and information sets of decision-makers will poorly screen out explanations that are merely possible given a set of information. Indeed, the researcher's efforts in source criticism, hermeneutics, and contextualization implies putting an explanation to a more severe test by requiring it to square with the historian's understanding of what they observe in the first place and the motivations and reasoning of decisions focal actors actually made. Thus, historical methods are intended to achieve a particular kind of *consilience:* an informed attempt at identifying and observing facts that would be observed if and only if a particular theory is true. Note that if-and-only-if requires exhaustive consideration of alternative explanations. Since many of these will be outside an initial explanatory conception, the process is necessarily iterative and will require weighing whether particular contextual details are or are not likely given different explanations.

Up until this point, we've ignored the central fault in the First Leg. Learning from past events to inform strategic decisions requires more than establishing the most likely explanations of past strategic decisions. We now consider this fault head on since we wish to use our learnings to inform today's strategists! Using additional historical methods to help augment difficult to interpret statistics is insufficient if those statistics are not inherently of interest. How can we use this information if the L.A.T.E. effect is of little relevance? How do we repair the first broken leg?

The *lovely* in abduction implies that an explanation provides meaning. And meaning requires that we consider the purpose of the explanation. The first stool is broken because a *likely* explanation for an average effect is not *lovely* because it provides little meaning to the strategic decision-maker. A

very detailed, severely tested causal theory may be of little general interest outside a setting.[11] Moreover, if our goal is to provide useful prescriptions based on causal mechanisms we expect to operate in novel settings, we will wish to seek out lessons from our studies that we can apply elsewhere. In the parlance of abduction, explanations that provide such meaning are *lovely*.

In our case, the most meaning for either a strategist, or a fellow scholar, will be an explanation that is **useful** in strategic design. Fortunately, we have some tools to help characterize an explanation's virtues that will help the author and reader ascertain an explanation's *loveliness*. We have written about this elsewhere (Pillai, Goldfarb, and Kirsch 2024), so will aim for brevity here. First, all causal explanations are theoretical, and if much of our theory relies on premises that have already been severely tested, then we will be more likely to believe an explanation to be true apriori. Such explanations have the virtue of *coherence*. Often, we refer to high coherence as "strong priors". Thus, *coherence* directly affects our assessment of *likely*. However, since most theoretical building blocks have not been severely tested in any setting, or if the *generality* of the theory is contested, then scholars will disagree on *coherence*. This statement is no more profound than pointing out that priors are subjective. But strategic design requires reducing the complexity of problems. So *parsimony*, sometimes labeled *simplicity* must be important. Simpler explanations are powerful, as they are more likely to transcend a setting. However, often, *parsimony* leads to more incomplete explanations (this is explicit in statistics such as adjusted R-squared or Bayesian Information Criteria).[12] However, we may want more than that! We sometimes want explanatory *depth* or *precision*, that is, a more detailed understanding of a causal mechanism. But this is problematic, because greater depth implies more assumptions, and links - many will be difficult to severely test - and hence greater depth may imply an explanation that is less *likely*. However, the greater depth will also allow more opportunities for severe testing. Additionally, greater *depth* may make an explanation more specific to a setting, and hence compromise *generalizability* (which had already been thrown into question when we recognized the broken first leg). And while this may be true, an explanation that has great depth also may still be quite *lovely*. The very fact that it is so deeply embedded in a context so as to specify complex linkages

[11] To be clear, if our goal is to document the truth - then likely is all we care about. But it is not. Our theories are incomplete and leave much of the variation in outcomes unexplained. We require this parsimony to sensemake. It may seem like a radical suggestion that we as scholars are interested in anything but the most *likely* explanations. But this cannot be the case! To see this consider that human cognition is fundamentally incapable of seeing *all* of reality. Instead, we simplify our explanations of events and choose to ignore some factors, and focus on others. (We label such factors as "outside the theory", or "idiosyncratic", and throw them in the error term). We do this because without parsimony, it is very difficult to make sense of what we see at all.

[12] It should be clear that these epistemic virtues related, with unclear degrees of orthogonality and their assessments are incommensurate judgments.

may allow the decision maker greater dexterity in figuring out which linkages may be applied to new settings. This dexterity comes from understanding past causal linkages in depth, understanding current settings in depth, and then using theory to choose which of those linkages from the past may be operating in the present.

So where does that leave us? A particular study will produce claims that have been tested with varying degrees of severity. These claims will rely on a series of judgements, and hopefully, the author will have transparently described not only the degree of severity to which an explanation was tested, but also their preferences for specific explanatory virtues, specifically explanatory *depth*, *coherence, parsimony,* and *generalizability*. With the clarification of these virtues, the author's judgments are better explicated, and hence the reader better equipped to assess the author's claims.

But repairing the first leg is more difficult. The idiosyncratic limitations of our sample is an insidious methodological problem. Mayo (2018) considers a similar problem. Most theories of importance are not refutable in any general sense. By this measure, Einstein's Theory of Relativity, or Darwin's Evolution, are not "scientific", at least in the Popperian sense. But we can examine aspects, assumptions, links that we expect to be in a causal chain. We can severely test bits and pieces, this implication or that. Perhaps we can assemble a series of "reusable microlinks" of causal sequences that we might expect to find in the causal chains that come from many strategic decisions. This relegates causal inference to something more modest - a tool that provides inputs into our thinking about complex, dependent, idiosyncratic, idiosyncratic strategic problems. Algorithmic solutions will ever elude us, but we can short circuit the building of expertise by severely testing some causal links in particular settings, and then equipping the strategist with a set of tools to ascertain to what degree they may apply in a new strategic landscape. Causal inference may help us choose hypotheses to allow the design of strategies, or inform experiments to develop strategies in frameworks such as, say, "Bayesian Entrepreneurship" (Ajay Agrawal, Arnaldo Camuffo, Alfonso Gambardella, Joshua S Gans, Erin L Scott, Scott Stern 2024).

What do we mean by "reusable microlinks"? These are the bits and pieces of causal chains. For example, we might be able to causally identify a link between whether an audience is aware of the prominence of any author and their consideration of an author's ideas (Simcoe and Waguespack 2011) - at least in the context of open software standards development. This may give us a clue as to how to interpret a new paper's success in Strategic Management. We have a plausible microlink reused from Tim and Dave's Big "et al." Adventure. However, we will need considerable judgment

then to design a strategy around this building block, even if we *judge* the link to be a candidate micro-link - potentially reusable in the strategic management context.

Thus, the first leg of the stool is reconstructed. We can provide characterizations of the severity of our tests of causal links in particular settings. Due to the required contextual depth necessary to do this well, we can weight it based on our judgment of the degree in which it was severely tested, and consider carefully the context of those explanations, and then use that to understand when and where causal chain micro-links from a particular study might be applicable to a new setting.

Conclusion

The journey of understanding causal inference in the realm of strategy has led us to recognize the limitations and potential of our current methodologies. Traditional statistical approaches, with their roots in frequentist statistics, often falter when applied to the unique and complex decisions that define strategic management. This is largely due to the inherent contextual and interdependent nature of strategic decisions, which make generalizable causal claims difficult to establish and often unreliable.

The first leg of our causal inference stool, the claim that strategic questions can be answered through statistical methods, is fundamentally broken. Strategic decisions are inherently unique, reflecting the specific context in which they are made. This uniqueness challenges the generalizability that statistical methods seek. The second leg, the reliance on pre-specification to ensure the reliability of causal claims, is similarly flawed. The flexibility required in strategic research often necessitates interaction with the data, which in turn compromises the reliability of frequentist statistical tests. The third leg, the dependency on knowing the true data-generating process (DGP), also crumbles under scrutiny. The iterative process of model specification and data interaction complicates the interpretation of statistical tests and undermines our ability to evaluate if the identified mechanisms genuinely fit the data.

Given these challenges, a shift towards a more nuanced approach is necessary. Inference to the Best Explanation (IBE), grounded in abduction, offers a promising alternative. This approach does not seek to provide definitive causal claims applicable across contexts but rather aims to construct plausible explanations for specific observations. Abduction considers that explanations are both lovely and likely. The rigor of causal inference steers us towards leveraging the ideas of severe testing, as proposed by Mayo and Cox, to rigorously evaluate hypotheses within the constraints of

our data and context. Severe testing emphasizes the need for a hypothesis to pass rigorous and fair tests that it would likely fail if false, thus offering a robust framework for causal inference. However, given the nature of strategic decisions, and the limitations of statistics, we should expect it challenging to severely test many causal mechanisms of interest. Thus, we need to heavily rely on contextual information outside our datasets and clarify the virtues of our explanations, and their ranking. The Best in Inference to the Best Explanation is a judgment - the author need make these judgments clear.

While this approach does not completely resolve the limitations of causal inference in strategy, it provides a structured way to navigate them. It allows us to build a more modest but practical stool, where the broken pieces of traditional methods are repurposed to support localized, context-specific explanations. These explanations are not broad generalizations but detailed accounts that can inform strategic decisions in specific settings.

The process of abduction, integrated with severe testing, requires an iterative and transparent methodology. Researchers must consider all plausible explanations and subject them to rigorous testing. This process acknowledges the trade-offs between model precision and generalizability, balancing the depth of understanding with the practical need for actionable insights. It also emphasizes the importance of triangulating multiple sources of evidence to build robust explanations.

In practice, this means that strategy researchers should embrace the iterative nature of explanatory calibration, where initial hypotheses are continually refined based on new data and insights. This iterative process aligns with the reality of strategic decision-making, where managers must adapt to evolving circumstances and refine their strategies accordingly.

Furthermore, the concept of reusable microlinks, or bits and pieces of causal sequences identified in specific contexts, offers a practical tool for strategists. These microlinks, once severely tested in particular settings, can provide valuable insights and inform strategic decisions in new contexts. While they do not offer a one-size-fits-all solution, they equip strategists with a repertoire of validated causal links that can be judiciously applied to novel situations.

In conclusion, while traditional causal inference methods face significant challenges in the field of strategy, the integration of IBE and severe testing offers a viable path forward. This approach embraces the complexity and context-dependence of strategic decisions, providing a framework for constructing robust, localized explanations. By focusing on severe testing and the iterative refinement of hypotheses, strategy researchers can navigate the limitations of statistical

methods and contribute valuable insights to the field. This modest yet practical stool, built from the broken pieces of traditional methods, holds the promise of advancing our understanding of strategic causal inference and enhancing the effectiveness of strategic decision-making.

References

Ajay Agrawal, Arnaldo Camuffo, Alfonso Gambardella, Joshua S Gans, Erin L Scott, Scott Stern. 2024. "Bayesian Entrepreneurship."

Bettis, Richard A. 2012. "The Search for Asterisks: Compromised Statistical Tests and Flawed Theories." *Strategic Management Journal* 33 (1): 108–13.

Bettis, Richard A., and Daniela Blettner. 2020. "Strategic Reality Today: Extraordinary Past Success, but Difficult Challenges Loom." *Strategic Management Review* 1 (1): 75–101.

Douven, Igor. 2022. *The Art of Abduction*. MIT Press.

Frake, Justin, Anthony Gibbs, Brent Goldfarb, Takuya Hiraiwa, Evan Starr, and Shotaro Yamaguchi. 2023. "From Perfect to Practical: Partial Identification Methods for Causal Inference in Strategic Management Research." *Available at SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4228655.

Goldfarb, B., and A. A. King. 2016. "Scientific Apophenia in Strategic Management Research: Significance Tests & Mistaken Inference." *Strategic Management Journal*. https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.2459.

Guzman, Jorge. 2022. "Treatment Effects in Managerial Strategies." SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3915606.

King, A. A. 2023. "Writing a Useful Empirical Journal Article." *Journal of Management Scientific Reports*. https://journals.sagepub.com/doi/abs/10.1177/27550311231187068.

King, A., B. Goldfarb, and T. Simcoe. 2021. "Learning from Testimony on Quantitative Research in Management." *Academy of Management Review. Academy of Management*. https://journals.aom.org/doi/abs/10.5465/amr.2018.0421.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *The American Economic Review* 73 (1): 31–43.

Lee, Gwendolyn, and Michael D. Ryall. 2024. "Duty to Elaborate One's Causal Theory: Toward a New Norm for Empirical Reporting in Strategy."

Leiblein, M. J., J. J. Reuer, and T. Zenger. 2018. "What Makes a Decision Strategic?" *Strategy Science*. https://pubsonline.informs.org/doi/abs/10.1287/stsc.2018.0074.

Lewbel, Arthur. 2019. "The Identification Zoo: Meanings of Identification in Econometrics." *Journal of Economic Literature* 57 (4): 835–903.

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press.

Mayo, Deborah G., and D. R. Cox. 2006. "Frequentist Statistics as a Theory of Inductive Inference." In *Optimality*, 77–97. Institute of Mathematical Statistics.

Pillai, Sandeep Devanatha, Brent Goldfarb, David A. Kirsch, and Evan Starr. 2024. "From Hypothesis Testing towards Inference to Best Explanation: Testimonial Structure for Abductive Studies in Strategy."

Pillai, S. D., B. Goldfarb, and D. Kirsch. 2024. "Lovely and Likely: Using Historical Methods to

Improve Inference to the Best Explanation in Strategy." *Strategic Management Journal*. https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.3593.

Pillai, S. D., B. Goldfarb, and D. A. Kirsch. 2020. "The Origins of Firm Strategy: Learning by Economic Experimentation and Strategic Pivots in the Early Automobile Industry." *Strategic Management Journal*. https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.3102.

Rumelt, Richard. 2011. *Good Strategy, Bad Strategy*. Crown Business.

Simcoe, Timothy S., and Dave M. Waguespack. 2011. "Status, Quality, and Attention: What's in a (Missing) Name?" *Management Science* 57 (2): 274–90.