SYMPOSIUM

All Else Equal in Theory and Data (Big or Small)

Scott Ashworth, *University of Chicago*
Christopher R. Berry, *University of Chicago*
Ethan Bueno de Mesquita, *University of Chicago*

The forms of explanation that dominate political science research in the formal-theory and causal-inference traditions are closely connected. Specifically, each explanation makes essential use of different but related types of all-else-equal claims. The emergence of "big data" already has begun to alter the landscape of empirical social science by making new sources of information (and, thus, new phenomena) amenable to quantitative analysis. However, neither the value of explanations rooted in all-else-equal claims nor the challenges associated with providing them will be altered in the least by big data.

ALL ELSE EQUAL (ON AVERAGE)

The formal-theory and causal-inference traditions in political science share deep connections. Both approaches strive to learn about mechanisms that play a role in explaining social phenomena, and each relies on all-else-equal comparisons to explore these mechanisms. To better understand the connections, we begin by outlining our view of how each approach works. We then explicate a bridge between them.

Formal, game-theoretic models are simplifications, meant to represent a small piece of a larger phenomenon.[1] The goal of such models typically is to offer qualitative insight into the mechanisms of a social process or pattern of behavior. This goal dictates simple models, for at least three reasons.

First, when we say that a model offers insight into mechanisms, we mean that it makes clear which assumptions generate a particular outcome in the model. If some of those assumptions and outcomes, and the relationships among them, represent analogous features of the empirical political environment, then the model is useful for explanation. Without knowing which features of a model drive a result, we cannot know whether the proposed explanation is robust or the product of special assumptions that are too distant from reality to constitute a satisfying explanation. This type of clarity typically is achievable only in a simple model. Perhaps the most important reason is that it is easier to "see through" models that do not have too many moving parts. Another reason is that abstraction and simplification also highlight robust categories of well-understood incentive problems—signaling, multitasking, externalities, bargaining, commitment problems, coordination problems, and so on—that describe numerous social situations and support explanation by analogy.

Second, theorists often want to ask what would happen if some feature of the environment changed. They answer these questions by asking what happens when the representation of that feature changes in the model. In this way, models can generate various types of insights, including empirical hypotheses, theoretical explanations, and normative implications. However, this process requires a model that can be manipulated, which necessitates simplicity. Morgan (2012) emphasized this aspect of formal models.

Third, applied-game theorists typically want more than a model in which agents behave in a way consistent with behavior in the world. Rather, they seek an understanding of agents' behavior in their models in terms of preferences, beliefs, information, and constraints. Thus, another important value of simple models—the analyst's inability to see through a more complex model thwarts hermeneutic understanding.

The fundamental explanatory goals of formal theory require simple models. However, this simplicity constrains the type of conclusions that we can draw from a formal model. These models typically are not attempts to write down a complete description of the mechanisms underlying some phenomenon. And since a theorist deliberately omits from any particular model most of what is important about the world, conclusions drawn from the model are—of necessity (implicitly or explicitly)—all-else-equal claims. For instance, a theoretical model's comparative static prediction about the relationship between two variables explicitly holds constant everything else in the model. The model also implicitly holds constant everything else that is omitted from it but might alter the relationship in the world.

We now turn to the causal-inference tradition of empirical research. Obviously, the goal of such work is to estimate causal effects. What are these effects? A general definition comes from the potential-outcomes framework (Holland 1986): the effect of x on y, for a particular unit, is the difference in the values of y for different assignments of x. Because this effect is defined at the level of an individual unit, it is an all-else-equal relationship.

A fundamental challenge for empirical researchers is that it is impossible, in principle, to simultaneously observe the same unit under two different treatment statuses. Hence, in a deep sense, it is impossible to directly observe a causal effect. In practice, empirical researchers try to hold all else orthogonal—that is, all else equal on average. The "gold standard" for doing

doi:10.1017/S1049096514001802 © American Political Science Association, 2015
https://doi.org/10.1017/S1049096514001802 Published online by Cambridge University Press

so is experimental randomization. Often, practical or ethical considerations render this impossible; therefore, empiricists interested in causal inference must find research designs that allow them to credibly claim to have held all else equal, on average, in observational data.

It is clear now why there is a tight connection between the formal-theory and causal-inference research traditions. Formal models, of necessity, generate all-else-equal claims. Moreover, because formal models omit so much that is important at the individual level, these claims are best thought of as all-else-equal claim about the central tendency of behavior in collections of individuals with preferences, beliefs, information, and constraints similar to those in the model. The causal-inference tradition is focused on research designs that make credible the claim to have held all else equal, on average, across these groups. Thus, the two traditions are naturally complementary.

This view of the role of theory and its connection to empirical analysis suggests rethinking some of the ways that political scientists typically conceptualize this relationship. We highlight two of those ways.

First, political scientists often speak of statistical models as though the goal were to estimate the entire process in the world that generates some outcome (sometimes casually referred to as the "data-generating process"). It is natural, from this point of view, to think that the goal of empirically useful theory is to explicate these processes in their entirety. Combining these ideas leads to the view that the variables included in the statistical and the formal models should be the same—that is, the formal model dictates what should be on the right-hand side of the regression.

Our view is quite different. Formal models generate qualitative hypotheses about particular all-else-equal relationships; the goal of empirical work motivated by this theory is to credibly assess those claims.

Our view is quite different. Formal models generate qualitative hypotheses about particular all-else-equal relationships; the goal of empirical work motivated by theory is to credibly assess those claims. Neither approach is about describing the entire process that gives rise to the outcome. Hence, the standard for whether a variable should be included in a regression is neither whether it was included in the motivating formal model nor whether that variable is part of the process that generates the outcome. Rather, the standard is whether omitting the variable would confound estimation of the relevant all-else-equal relationship.

Second, viewing all-else-equal claims as a bridge between theoretical and empirical research points to a way that the approaches developed by causal-inference scholars are actually of broader applicability than their name suggests. After all, the all-else-equal claims that come from formal models need not be causal (although they often are). Empirically assessing a non-causal but still all-else-equal theoretical claim requires the same rigorous attention to confounding that is required for estimating causal effects. As such, the research designs developed for causal inference are often the "right tool for the job"—even when the job is not causal inference.

We can illustrate both points in an example about the performance of men versus women in the US Congress (see Anzia and Berry 2011). Consider a simple setting in which an electorate must choose between two candidates, each drawn randomly from the same talent pool. The voters prefer more talented candidates but they also discriminate against women. If faced with a male and a female candidate, the voters will elect the female candidate only if she is sufficiently more talented to overcome the voters' bias against women. (It is straightforward to formalize this model; however, in the interest of conserving space, we do not.)

This theoretical model makes a causal prediction: increased voter bias causes an increase in the average quality of elected women and a decrease in the average quality of elected men. However, the theoretical model also makes an interesting non-causal prediction: if talent translates into performance, then, all else equal, elected women will perform better than elected men, on average. This prediction is non-causal because the theoretical model does not assert that turning an elected official from a man into a woman improves performance; nevertheless, it is an all-else-equal claim.[3]

Empirically assessing this non-causal prediction requires as much attention to issues of confounding as for the estimation of a causal relationship. Imagine a researcher taking this prediction to data from the US House of Representatives to compare how well male and female representatives perform at procuring federal spending for their districts. Suppose the researcher found that women perform better, on average. The researcher would be concerned that this difference might reflect attributes of the districts—for example, they are poorer, have larger African American populations, and are more urban—that are associated with electing women and also lead those districts to attract more federal spending. As such, to credibly claim to have provided evidence for the theoretical model's explanation, the researcher would need a research design that credibly held all else equal, even though the prediction being investigated is not causal.

Moreover, these concerns about specific confounders did not come from the theoretical model. That model is not (and was not intended to be) a complete description of the process by which federal funds are allocated. It is a simplification intended to highlight the impact of voter discrimination. As such, it would have been an absurd distraction for the theoretical model to include district race and urbanness—which, in fact, are correlated with federal spending and legislator gender but were not the subject of the theoretical analysis. That the theoretical model does not suggest district characteristics as

https://doi.org/10.1017/S1049096514001802 Published online by Cambridge University Press

confounders, however, does not justify the empirical researcher omitting them. The theoretical model offers an all-else-equal prediction. The job of the empirical researcher, if she wants to assess this prediction, is to find a research design that takes seriously this all-else-equal condition by addressing all possible confounders, whether or not they appear in the theoretical model.[4]

DATAFICATION

The term "big data" emerged in computer science to describe problems associated with datasets too large to be read into the memory of a single computer (Cox and Ellsworth 1997). In this sense, big data remains an active area of research in computer science (National Research Council 2013). However, this is not what we are concerned with in this paper. In political science, the term has come to refer, loosely, to a collection of techniques associated with data mining and machine learning.[5] (See, e.g., the *Political Analysis* virtual issue, "Big Data in Political Science.")

The tools of machine learning have proven productive in the social sciences by automating various aspects of data collection, a process we call "datafication." The most straightforward version of datafication is collecting data from many widely distributed locations, as in web scraping. In addition to data collection, machine-learning techniques can be used to convert nontraditional sources of information—for example, text and images—into data that can be used for statistical analysis. For instance, Grimmer (2010) used unsupervised learning techniques to code the substantive content of congressional press releases in terms of the policy agendas that they prioritize. It is impractical to do this data collection manually; therefore, it was only with the advent of machine-learning techniques and powerful computers that these information sources became accessible to quantitative researchers.

It is critical to distinguish between the use of machine learning for datafication and for inference-oriented data analysis. To understand this distinction, consider Gentzkow and Shapiro's (2010) analysis of the extent to which media slant is driven by consumer demand versus the ideology of newspaper owners. They used machine-learning techniques for text categorization to classify the partisanship of newspapers. In particular, they measured media slant by comparing the words used in newspaper editorials to those used in speeches by congressional Republicans and Democrats. After this datafication was accomplished, Gentzkow and Shapiro addressed their substantive question with a combination of theory and traditional causal-inference approaches. They first showed that observed media slant is inconsistent with the predictions of a theoretical model in which variation in slant is driven only by demand-side factors. Given this, they turned to more direct attempts to decompose the effects of consumer and owner ideology: they compared the media slant of newspapers with the same owner located in markets with different demand-side characteristics and they used religiosity to provide plausibly exogenous variation in consumer political attitudes.

In our view, Gentzkow and Shapiro's (2010) study exemplifies the promise of big data for the social sciences. Datafication unlocked new information for quantitative analysis; however, those data were analyzed in light of a research design that exploited plausibly exogenous sources of variation in consumer political attitudes. Without such a research design, it would have been impossible to provide a credible answer to the research question—that is, all-else-equal, what are the separate effects of consumer and publisher ideology on media slant? A naïve regression of newspaper slant on consumer ideology would have done nothing to address the obvious problems of unobservable confounders—for instance, that conservative newspaper owners simply might be more likely to locate in conservative markets. This is because the naïve regression depends on selection on observables as its research design—an incredible assumption in this setting. It follows that any method that relied on selection on observables for identification (e.g., matching) would not have yielded credible estimates of the causal effect. Indeed, the same also is true of machine-learning techniques that assume selection on observables—including the most sophisticated techniques for estimating treatment effects, such as the post-double-LASSO estimator of Belloni, Chernozhukov, and Hansen (2014).

The more breathless accounts of "big data" might lead us to suppose that this task of finding credible research designs also can be automated. But this can't be right. Identifying a source of exogenous variation requires creative insight about the phenomenon under study. Arguments for the credibility of such a source of exogenous variation rest on substantive knowledge of that phenomenon. No existing computational technique can come remotely close to doing either.[6] For the time being, good empirical research will continue to rely on actual people doing actual thinking about social phenomena.

FAMILIAR MISTAKES AT SCALE

Not only do machine-learning techniques not help with the development of a research design, in fact they also can be harmful for inference, even in the presence of a research design.

To understand the issue, imagine that we have a perfect research design: experimental randomization of a treatment. At least two things could go wrong. First, suppose there is only one outcome but the researcher observes many covariates regarding the subjects. In principle, the researcher can search across all of those covariates and all of their possible interactions to find a subpopulation on which the treatment effect is statistically significant. Of course, given enough independently varying covariates, such a subpopulation is guaranteed to exist, even if there is no real treatment effect.

Second, suppose there are many possible outcomes that, in principle, could have been affected by the treatment. Again, the researcher can search across all of those outcomes to find a statistically significant treatment effect. Given that such effects will be found spuriously 5% of the time, the researcher who conducted 100 such tests should expect to find five significant effects, even when there are no real treatment effects. When only the significant results are reported, the reader has no basis for judging the probability that the results in question arose by chance—even when that probability may be much closer to 1 than to 0.05.

https://doi.org/10.1017/S1049096514001802 Published online by Cambridge University Press

Of course, the problems of multiple-hypothesis testing did not originate with machine learning. Other than time constraints, nothing prevents a researcher from manually running all of the regressions just described. However, because machine learning automates and accelerates the process, we worry that it facilitates the industrial-scale production of spurious results through multiple hypothesis testing. When coupled with the well-known problems of publication bias (Gerber and Malhotra 2008), industrial-scale multiple-hypothesis testing may seriously retard and distort the accumulation of knowledge.

For the time being, good empirical research will continue to rely on actual people doing actual thinking about social phenomena.

One approach to the problem of multiple-hypothesis testing is technical—for instance, by attempting to adjust p-values or to control the false-discovery rate. Political scientists have much to learn from the work being done in computer science and statistics on this front. However, we think a narrow technical focus will miss a larger problem. As discussed previously, the goal of theoretically motivated empirical research is to estimate a particular all-else-equal relationship. Analysts should have a research design for the purpose of answering a specific research question. With this clear and guiding purpose, they have no reason to be searching across many specifications (or subpopulations or outcomes) for results. Indeed, to the extent that analysts find themselves searching across specifications, it should be to probe the robustness of a theoretically interesting relationship. Loosely speaking, analysts should be searching to ensure there are no reasonable specifications that make the theoretically motivated result disappear, rather than searching for the one specification that produces a statistically significant "finding."[7] In this sense, we see the nascent preregistration movement—which urges researchers to publicly state their research design and model specification before analyzing their data—and the associated willingness to publish null results as prerequisites for big data or machine learning to be truly useful contributors to inference-oriented empirical research in political science.

FORECASTING, INVARIANCE, AND POLICY

A more radical view is that the rise of big-data approaches will diminish the importance of explanation rooted in all-else-equal claims of the type discussed in this article. In that discussion, we argued that all-else-equal claims are crucial for scientific understanding. However, all-else-equal claims are also important as facilitators of mechanism-based forecasting of the effects of policy interventions. Machine-learning techniques complement a different approach to forecasting—allowing for purely statistical forecasting on a larger scale and with greater accuracy than previously could be conceived. Perhaps, then, big-data approaches will allow policy makers to substitute purely statistical forecasting for mechanism-based forecasting.

What do we mean when we distinguish between purely statistical and mechanism-based forecasting? At a conceptual level, we can understand the distinction by simply thinking about causal effects (as a special case of a mechanism) and ordinary least squares regressions. Suppose that there are two variables, y and x, and that the true causal effect of a change in x on y is given by β. Furthermore, suppose that there is another variable, z, which is correlated with both y and x. If a researcher regresses y on x but cannot observe z, then the omitted variables bias formula indicates that the regression coefficient, α, is equal to β, plus a correction term that depends on the correlation between x and z and partial correlation between y and z conditional on x. Purely statistical forecasting predicts y from x using α. The causal effect (β) is of no particular interest. The predictions of a mechanism-based forecast attempt to use information about β.

When do we want to take a purely statistical versus mechanism-based approach to forecasting? The advantage of the purely statistical approach is that it makes optimal use of all the information. In our example, it is precisely because α exploits correlations with z that it is statistically optimal—that is, the coefficient of the best linear predictor of y given x. Therefore, if we are interested in simply forecasting y based on the x's that are already out there in the world, then we typically want to estimate α.

However, sometimes there are costs to this purely statistical approach. Any forecast relies on some type of extrapolation. Concerns about forecasts often reflect skepticism regarding invariance through the extrapolation. Hume's problem of induction arises from concerns about invariance across time. External validity problems arise from concerns about invariance of treatment effects across sites. The possibility of non-monotonic dose-response relationships raises concerns about invariance of incremental treatment effects across treatment intensities. Lack of overlap raises concerns about invariance of treatment effects across covariate values.

The cost of purely statistical forecasting is that it relies on a strong invariance assumption that is typically false for policy interventions. Suppose that we are interested in forecasting the effects on y of a policy intervention that changes x. Having changed the relationship between x and z means that it no longer makes sense to think that the observed values of x are informative about z in the way that they were prior to the intervention. Forecasting as if the relationship between x and z were invariant, therefore, is a mistake.

To make this assertion more concrete, we consider a simple example. Suppose that we observe data on vote share and challenger campaign spending. It is plausible that it is easier for high-quality challengers to raise money and that they do better in elections for reasons independent of that money. The coefficient in a regression of vote share on challenger

https://doi.org/10.1017/S1049096514001802 Published online by Cambridge University Press

spending (α) reflects information about candidate quality as well as the true causal effect of spending (β). As such, it is a biased estimate of the causal effect of spending. Nonetheless, if we simply want to forecast election outcomes based on observed campaign spending, we need to use this regression coefficient because all of the information it contains is useful for the forecast. The situation is very different if we want to evaluate the consequences of public financing of campaigns for electoral competitiveness. A strong version of such a reform (e.g., one that simply gives all candidates the same funds) will eliminate the relationship between spending and quality. Hence, it would be a mistake to make a purely statistical forecast—that is, one that assumes the relationship between challenger spending and election outcomes will continue to be given by α after the reform.

Of course, the mechanism-based forecasting approach also relies on invariance assumptions to forecast the effects of this reform. Assume that we have a research design that identifies β. Now, if we simply want to forecast the portion of the effect of the reform that operates through the direct causal effect of spending, we need to assume that voter responses to challenger spending are invariant to the source of challenger funds. If we want to go further, and forecast the total effect of the reform, we also must assume invariance of all nonspending characteristics of challengers. Although both assumptions are arguably quite strong, they are weaker than needed for the purely statistical forecasting approach. In particular, if either of these assumptions does not hold, then (except for a miracle) α will not remain invariant. But the reverse is not the case.

The deep challenge of forecasting the effect of policy interventions is precisely that even invariance assumptions like these might well not hold. For example, invariance of voter response would not hold if the voters draw different inferences about candidates when they observe spending raised from donors rather than given from the public coffers (Ashworth 2006; Houser and Stratmann 2008; Prat 2002). Similarly, invariance of nonspending characteristics of challengers would not hold if the availability of public financing encourages a different group of candidates to seek office (Tillmann 2014). Each of these is a more general type of invariance failure—strategies of actors may not be invariant across policy regimes. (In some economic contexts, this point is called "the Lucas critique," after the forceful articulation of the point made in Lucas 1976.)

Typically, political scientists are called on to forecast the effects of policy interventions in settings—for example, campaign-finance regulation, civil-service reform, and counterinsurgency strategies—in which strategic adjustment is of the essence. In such settings, we cannot escape the Lucas critique. The data (whether big or small) are always drawn from the world that exists in the absence of the policy intervention. However, an accurate forecast of the effect of the intervention requires thinking counterfactually about situations for which in fact there may be no data (even in principle) to reveal how strategic adjustment will occur. In such settings, the only hope is a combination of theoretical accounts of mechanisms and data that are informative about the magnitude of the effects associated with those mechanisms that, together, might give the researcher some leverage over the key counterfactuals.

...an accurate forecast of the effect of the intervention requires thinking counterfactually about situations for which in fact there may be no data (even in principle) to reveal how strategic adjustment will occur.

ACKNOWLEDGMENTS

We benefited from conversations with Anthony Fowler, Pablo Montagnes, Rocío Titiunik, Dustin Tingley, and Stephan Wolton. We appreciate comments from the symposium guest editors, Bill Clark and Matt Golder ■

NOTES

1. Our thinking about the role of models in representation and explanation, and their relationship to empirical research, is deeply influenced by Hacking (1983) and Giere (1988).
2. Of course, "all" should not be taken literally. For instance, when estimating causal effects, it is a mistake to hold posttreatment variables equal.
3. This lack of a causal interpretation is distinct from the old issue of the potential conceptual incoherence of the idea that gender has a causal effect. We could describe the model and results in terms of voter discrimination about any feature of a candidate, including those that, in principle, are experimentally manipulable. Gender happens to be the subject of Anzia and Berry (2011).
4. Anzia and Berry (2011) used a difference-in-differences design.
5. This includes methods such as clustering, classification and regression trees, and penalized regression techniques such as the LASSO. Hastie, Tibshirani, and Friedman (2009) provide a canonical introduction to these methods.
6. We might hope that with enough covariates, selection on observables will become a credible research design for many problems. We find it implausible that this will be the case in the short-to-medium term. See Titiunik (2015), in this issue, for elaboration on this point.
7. A related but distinct way to probe robustness involves fixing a specification and then finding bounds on the estimate, given a range of violations of the identifying assumptions. Rosenbaum's (2010) approach considers violations that are substantively "plausible," whereas Manski's (2007) approach considers all possible violations.

REFERENCES

Anzia, Sarah F., and Christopher R. Berry. 2011. "The Jackie (and Jill) Robinson Effect: Why Do Congresswomen Outperform Congressmen?" *American Journal of Political Science* 55 (3): 478–93.

Ashworth, Scott. 2006. "Campaign Finance and Voter Welfare with Entrenched Incumbents." *American Political Science Review* 100 (1): 55–68.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies* 81 (2): 608–50.

"Big Data in Political Science" *Political Analysis* Virtual Issue, http://www.oxfordjournals.org/our_journals/polana/virtualissue4.html.

Cox, Michael, and David Ellsworth. 1997. "Application-Controlled Demand Paging for Out-of-Core Visualization." In *Proceedings of the 8th Conference on Visualization '97 (VIS '97)*. 235–ff. Los Alamitos, CA: IEEE Computer Society Press. Available at http://dl.acm.org/citation.cfm?id=266989.267068.

https://doi.org/10.1017/S1049096514001802 Published online by Cambridge University Press

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from US Daily Newspapers." *Econometrica* 78 (1): 35–71.

Gerber, Alan, and Neil Malhotra. 2008. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3 (3): 313–26.

Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18 (1): 1–35.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. New York: Cambridge University Press.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition. New York: Springer-Verlag.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Houser, Daniel, and Thomas Stratmann. 2008. "Selling Favors in the Lab: Experiments on Campaign Finance Reform." *Public Choice* 136 (1–2): 215–39.

Lucas, Robert E., Jr. 1976. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets*, Karl Brunner and Alan H. Meltzer (eds). Carnegie-Rochester Conference Series on Public Policy, Volume 1, 19–46. Amsterdam: North-Holland.

Manski, Charles F. 2007. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

Morgan, Mary. 2012. *The World in the Model: How Economists Work and Think*. New York: Cambridge University Press.

National Research Council. 2013. *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press.

Prat, Andrea. 2002. "Campaign Advertising and Voter Welfare." *Review of Economic Studies* 69 (4): 997–1017.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer.

Tillmann, Philipp. 2014. "Entry into Electoral Races and the Quality of Representation." University of Chicago manuscript.

Titiunik, Rocío. 2015. "Can Big Data Solve the Fundamental Problem of Causal Inference?" *PS: Political Science and Politics* 48 (1): this issue.

https://doi.org/10.1017/S1049096514001802 Published online by Cambridge University Press