SYMPOSIUM

# The Discipline of Identification

**Luke Keele,** *Pennsylvania State University*

This article is a commentary on whether there are tensions among big data, theory (mostly formal), and the statistics of causal inference. Of these three areas, big data is clearly the newest. However, the idea that political methodology should focus on causal inference also is a relatively new idea. I have attended the Annual Political Methodology Summer Meeting since 2000. At that first meeting and at several afterwards, there were no papers that focused on the statistics of causal inference; by 2009, causal inference was the conference theme.

The relatively new focus on causal inference contributes to some of the confusion about exactly what it means to focus on this topic. To many, causal inference is associated with particular techniques such as matching and regression-discontinuity designs. Although these methods are important, they are simply tools in service of a broader approach to statistical analysis. Given the confusion about what it means to focus on causal inference, this paper begins with an overview of the causal-inference approach to statistical analysis, which I argue has wide relevance—if only to clarify the goal of statistical analysis. Based on this summary, I outline how causal inference and what I term the "discipline of identification" contribute to the development of big-data approaches to political science. Finally, I discuss the interplay between theory and causal inference, in which I think conflicts can and do arise.

## CAUSAL INFERENCE

I was not trained in causal inference. My graduate training in methodology entailed various forms of regression analysis with a focus on time-series models. My dissertation and early methodology publications focused on time-series models. It was only later that I refocused my research on causal inference. Although the rise of causal inference closely parallels the rise of matching estimators, causal inference is not matching. I came to causal inference through matching; however, as I read the literature that accompanied many of those papers, I realized something about the general practice of statistics. In the applied use of statistics, researchers (including myself) often were far too lax about the role of assumptions in our analyses. Whereas assumptions were often presented in statistical texts, the implications of violating assumptions were ill-defined, and the question of whether the parameters in statistical methods (e.g., linear regression) could be causal effects was largely ignored. As I became familiar with the literature of matching—specifically, the statistics literature based on potential outcomes—it was clear to me that issues of causality were hidden in most of my analyses.

My study of this literature led me to understand that the role of assumptions is critical when statistical analysis is focused on causal questions. Why are assumptions important? Manski (2013) makes the following point:

$$\text{Data} + \text{Assumptions} = \text{Conclusions}$$

This is best illustrated by an example from Keele and Minozzi (2012). In that analysis, we studied whether Election Day registration (EDR) increased turnout in the states of Minnesota and Wisconsin. The outcome was a simple binary measure of whether or not a person voted and EDR served as a treatment that might increase turnout. What could we learn from data about the effect of EDR on turnout if we made no assumptions? Manski (1990) derived what are known as the "no-assumption bounds," which provided a range for the EDR treatment-effect estimates based on the data alone. These bounds revealed exactly how much we could learn without making any assumptions. Using data from 1980, the no-assumption bounds for the effect of EDR were -63 and 37. That is, EDR may have decreased turnout by as much as 63% or increased it by as much as 37%. Clearly, by using the data alone, we learned little about the effect of EDR.

Thus, assumptions are an all-important element in any attempt to infer causal relationships with data. In a causal-inference framework, the role of assumptions is defined through the concept of identification. Researchers want to observe causal relationships; however, data reveal only associations, which are a combination of causal and non-causal (i.e., spurious) components. Without assumptions, there is no way to separate the causal component of an observed association from the spurious components. In an identification analysis, we prove which assumptions are necessary to reduce an association to its causal component. When those assumptions hold, the causal effect is "identified." We may hear people question whether a statistical estimate is identified. They are asking whether the set of identification assumptions invoked is sufficiently plausible for an estimated statistical association to be given a causal interpretation. Identification is a completely separate task from estimation, which has been the more general focus in statistics.

This concept of causal identification is different from other concepts of identification in statistics, since the unobservable quantities are different. For example, many discrete-choice models are not identifiable unless we make assumptions about the parameters of unobserved probability distributions. With identification in causal inference, what is unobservable are counterfactuals related to treatment regimes.

The focus on identification gives greater clarity to the assumptions needed for statistical analyses to provide estimates of causal effects. It is this focus that has led to a greater emphasis on randomized experiments and regression-discontinuity designs, in which the assumptions needed to identify causal


© American Political Science Association, 2015
doi:10.1017/S1049096514001826
https://doi.org/10.1017/S1049096514001826 Published online by Cambridge University Press

effects are much weaker. The key threats to identification—confounding and selection—receive greater attention in the work on causal inference than typically has been the case. This deeper understanding of focus on identification has led to greater skepticism about what can be accomplished using statistical estimates, and it has pushed researchers to plausibly identify causal effects using a wider variety of strategies. This is the revolutionary part of the "identification" or "credibility" revolution (Angrist and Pischke 2010).

In general, the identification revolution has resulted in something I term the "discipline of identification" in which there is a different stance toward analyses that purport to be about causal effects. Analysts must articulate an identification strategy, defend why it is credible, and provide empirical evidence to show that either identification assumptions are plausible or inferences are invariant to those assumptions. This is in contrast to simply estimating a statistical model and interpreting the results, with little time spent discussing whether such effects are causal and what must be true for these effects to be given a causal interpretation. Here, identification provides discipline by requiring analysts to clearly define and defend the assumptions needed for a statistical analysis to be causal.

What happens if we do not have the discipline of identification? I use the regression model as a useful case study of what occurs without it. I use the term "regression model" loosely to encompass not only models estimated with least squares but also panel-data models and generalized-linear models, among a wide variety of these models. Much of the instruction on statistical analysis in political science focuses on teaching students how to estimate various types of regression models. Bivariate regression models clearly can be used descriptively to fit a line through data. In these cases, the estimate in a regression model is simply a measure of statistical association between X and Y using a linear equation as the summary. Thus, we can treat regression models as a non-causal method that provides a useful summary of the data.

However, most of the focus is on regression models with control variables. The logic for the shift to multivariate models is that such models are said to be unbiased if the analyst has justified them because identification holds and the regression model is estimating a causal effect. However, if the regression estimate is intended simply to represent the statistical association between X and Y, one could argue there is no need to estimate a multivariate model. This leads to a paradox in the use of regression models. We teach students about various types of regression models, and there is a statement of assumptions that would lead those models to be causal models. However, without a strong understanding of identification, there often is little emphasis on what those assumptions actually mean. Specifically, one key assumption is that we must control for all of the joint causes of X and Y. In a model of international conflict in which we want to estimate the effect of trade on conflict, we must observe and measure all of the joint causes of trade and conflict. This state of the world is not even approximately true, yet analysts regularly describe results from regression models of this type as the effect of X on Y.

A sound understanding of identification is what allows analysts to clearly distinguish between causal and descriptive statistical analyses. Other scholars in this dialogue make the case for description, and I also think it is an important task.

*A sound understanding of identification is what allows analysts to clearly distinguish between causal and descriptive statistical analyses.*

The difficulty is that without a fixed notion of identification, statistical analyses often occupy a "netherworld" somewhere between description and causal inference. Few analyses in political science remain purely descriptive, and many analyses do not clearly choose a side. The discipline of identification demands that instead of simply presenting model estimates and remaining silent on whether such parameters are descriptive or causal, analysts must clearly define whether the interpretation of that model is intended to be descriptive or causal. If the model is meant to be causal, then an identification strategy must be articulated and defended. Without the discipline of identification, it is too easy to confuse observed statistical associations for causal effects, it is too easy to confuse a descriptive analysis as a causal analysis, and it is too easy to use causal language to describe statistical analyses that are descriptive. The discipline of identification allows for a more lucid delineation between descriptive and causal analyses.

We might be concerned that the focus on causal identification and causal inference affects the types of empirical questions studied. Here, researchers might test only those theories that easily fit into the framework of, for example, the regression-discontinuity design and randomized experiment. This may be true, but it has always been true. For a long time, all tests of theory in political science resembled regression problems. It may be true that for some questions in political science, it will be difficult to achieve credible identification. However, only when we understand those limits, will researchers begin to think about new approaches that may produce better inferences.

## DATA, BIG OR OTHERWISE

In general, there is no tension between big data and causal inference. Identification abstracts away from the uncertainty caused by sample-size limitations; therefore, conclusions about identification are unrelated to sample size. Big data does not change the fact that conclusions are a function of both data and assumptions. The bias caused by confounding is present when there are one thousand or 100 million observations. In fact, big data can make these problems worse. In general,

https://doi.org/10.1017/S1049096514001826 Published online by Cambridge University Press

we tend to prefer larger datasets because our estimates will be more efficient. What is less well understood is how bias can grow as sample size increases with observational data. Cochran and Chambers (1965) proved that if there is a fixed bias, it will dominate the mean squared error of the estimate as the sample size grows.

An example puts this fact in context. Assume that we are interested in learning about the causal effect of education on political participation. Suppose that education and political participation are both caused in part by cognitive ability that we cannot measure. If this is true, cognitive ability will confound the education and political-participation effect. Any statistical association that we calculate between education and political participation will be at least if not entirely spurious because we cannot measure ability. As such, there is a fixed bias due to confounding, that bias from cognitive ability will dominate that statistical estimate as the sample size grows. Our standard errors will shrink as the sample size grows, but those errors reflect only sampling uncertainty; they indicate nothing about our uncertainty due to confounding. The result is that we have precisely estimated a highly biased effect in which the confidence intervals may no longer include the true value.

*In general, we tend to prefer larger datasets because our estimates will be more efficient. What is less well understood is how bias can grow as sample size increases with observational data.*

The difficulty is that the hype around big data has led to claims that we can abandon the discipline of identification because we have big data. An essay in *Wired* magazine outlines what many think big data can offer (Anderson 2008). One quote from the essay clearly ignores the insight about fixed biases: "Petabytes allow us to say: 'Correlation is enough.'" The problems caused by confounding and selection are still present regardless of how big the data is. In short, the discipline of identification is just as necessary with big data as it is with any size dataset. Although Anderson's essay rejects the need to think about identification, this view arises from fields such as computer science, in which statistical thinking is less common. In general, I think one strength of big data within political science is that statistical concepts are of central concern. For example, all of the representatives of big data in this symposium clearly recognize the importance of identification. To my knowledge, the exaggerated claims about big data are not made by researchers in political science. In fact, a number of techniques from the machine-learning methods used in big data find uses in the estimation of causal effects.

## THEORY

Strictly speaking, theory and causal inference do not conflict. Statistics are the tools by which we test theories, and statistical methods need guidance from theory. Causal theories are the precursor to causal analyses. I am not sure that anyone would disagree with that statement. Currently, I think novel theories are highly valued in the field; papers without them rarely are candidates for publication in the best political science journals.

Of course, there is nothing wrong with new theories; however, I believe that an overemphasis on theory can impede the ability of the discipline to establish causal relationships. One point of emphasis in the identification revolution is that causal inference is difficult. A series of regression models is hardly the last word in whether a causal theory holds. Within causal inference, it is generally understood that only a series of different studies using disparate research designs can provide solid evidence for a causal relationship. This understanding implies that, as opposed to a single test, a theory requires a number of tests from different research designs. We might assert that whereas theory is critically necessary, too many theories might be harmful. It is probably the case that the gain from an incomplete test of a new theory is less than the gain from a new test of an existing theory.

Economics provides a useful example. The question of whether attending college increases income is one of long-standing interest to economists. The theory behind the question is relatively simple. Although I hesitate to state that the causal hypothesis that college leads to higher incomes is definitively settled, a review of the many different research designs used makes a convincing case for a causal effect. If all of the researchers conducting these studies had been told by reviewers that new theory was needed, little progress would have been made.

When we overvalue novel theories, we tend to dismiss attempts to answer old questions with new research designs. We can value papers with new theories and little or no empirics. We also can value papers with little in the way of new theory but that present novel research designs that provide new empirical evidence about causal relationships. We could argue convincingly that both topics require such attention that it may be difficult to present both novel theory and empirics in the same paper. Currently, I would state that, in general, a paper that does not engage in new theory development will have a difficult time being published in a top journal. I don't think that is healthy. If we really want researchers to clearly establish causal relationships that is often worth doing alone in a single paper.

## CONCLUSION

Ultimately, this article is as self-serving as we would expect. I offer reasons for why my specialty is even more important than most people think. I argue that big data will go astray if attention is not given to the principles of identification.


https://doi.org/10.1017/S1049096514001826 Published online by Cambridge University Press

Finally, I argue that the discipline undervalues novel research designs, which comprise a key part of causal inference. These claims all may be true. At the very least, I hope readers understand that causal inference is about the basics of statistical analysis. Confounders are not going anywhere; it is only our attention that we often allow to be diverted. ■

## REFERENCES


Anderson, Chris. 2008. "The End of Theory." Available at http:// archive.wired.com/science/discoveries/magazine/16-07/pb_theory. Accessed September 3, 2014.

Angrist, Joshua D., and Jorn-Steffen 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.

Cochran, William G., and S. Paul Chambers. 1965. "The Planning of Observational Studies of Human Populations." *Journal of Royal Statistical Society*, Series A 128 (2): 234–65.

Keele, Luke J., and William Minozzi. 2012. "How Much Is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data." *Political Analysis* 21 (2): 193–216.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review Papers and Proceedings* 80 (2): 319–23.

———. 2013. *Public Policy in an Uncertain World: Analysis and Decisions*. Cambridge, MA: Harvard University Press.


https://doi.org/10.1017/S1049096514001826 Published online by Cambridge University Press

AMERICAN POLITICAL SCIENCE ASSOCIATION

# APSA Task Force Reports




**The APSA task force reports seek to expand the public presence of political science by:**

► applying political science research to public policy issues,

► and sharing exepertise on the most significant topics of public concern.

**APSA's current task force report!**

*Improving Public Perceptions of Political Science's Value*

Read the new APSA task force report, which identifies means by which individual scholars and professional organizations can make political science's insights and discoveries more accessible, more relevant, and more valuable to more people.

Many of the following reports are also available as **FREE** .pdf downloads!

Visit **www.apsanet.org/reports** to obtain your APSA task force publication.

apsa AMERICAN POLITICAL SCIENCE ASSOCIATION

For questions or concerns, please contact publications@apsanet.org.

https://doi.org/10.1017/S1049096514001826 Published online by Cambridge University Press