ANNUAL REVIEWS

Annual Review of Political Science

Testing Causal Theories with Learned Proxies

Dean Knox,[1] Christopher Lucas,[2] and Wendy K. Tam Cho[3]

[1]Operations, Information, and Decisions Department and Analytics at Wharton, The Wharton School of the University of Pennsylvania, Philadelphia, Pennsylvania, USA; email: dcknox@upenn.edu

[2]Department of Political Science and Division of Computational and Data, Washington University in St. Louis, St. Louis, Missouri, USA; email: christopher.lucas@wustl.edu

[3]Departments of Political Science, Statistics, Mathematics, Computer Science, and Asian American Studies; College of Law; and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA; email: wendycho@illinois.edu

Annu. Rev. Political Sci. 2022. 25:20.1–20.23

The *Annual Review of Political Science* is online at polisci.annualreviews.org

https://doi.org/10.1146/annurev-polisci-051120-111443

Keywords

causal inference, machine learning, supervised learning, measurement, proxies

Abstract

Social scientists commonly use computational models to estimate proxies of unobserved concepts, then incorporate these proxies into subsequent tests of their theories. The consequences of this practice, which occurs in over two-thirds of recent computational work in political science, are underappreciated. Imperfect proxies can reflect noise and contamination from other concepts, producing biased point estimates and standard errors. We demonstrate how analysts can use causal diagrams to articulate theoretical concepts and their relationships to estimated proxies, then apply straightforward rules to assess which conclusions are rigorously supportable. We formalize and extend common heuristics for "signing the bias"—a technique for reasoning about unobserved confounding—to scenarios with imperfect proxies. Using these tools, we demonstrate how, in often-encountered research settings, proxy-based analyses allow for valid tests for the existence and direction of theorized effects. We conclude with best-practice recommendations for the rapidly growing literature using learned proxies to test causal theories.

1. I DON'T KNOW Y (AND OTHER CHALLENGES ARISING FROM IMPERFECT PROXIES IN SOCIAL SCIENCE)

Social-scientific theories often involve latent concepts that are not directly observed by researchers, such as "democracy" or "ideology." To empirically evaluate their theories, researchers must imperfectly measure these unobserved concepts. Classic examples include the use of expert panels to rate countries' political systems and factor analysis to construct weighted indices from survey responses, which respectively produce proxies of democracy and ideology. While various forms of quantitative measurement emerged with the advent of empirical social science, the recent growth of machine learning has led to an increase in research that learns proxies using computational models. Compared to classic approaches—which often require costly in-depth expert reading or derivation of case-specific measurement models—this new body of work increasingly uses rich, unstructured data and flexible, off-the-shelf statistical tools to measure concepts of theoretical importance. In this article, we review common approaches and key methodological considerations in this rapidly growing literature. We focus specifically on best practices for incorporating imperfectly learned proxies into subsequent analyses, which pose underappreciated challenges for analysts seeking to rigorously test social-scientific theories. Excellent references are available for measurement (Adcock & Collier 2001) and statistical learning (Grimmer et al. 2021) more broadly. In contrast to these and similar review articles, we focus on the use of learned proxies in causal tests, particularly where the proxy is estimated from a computational model.

Regardless of how formally they are expressed, social-scientific theories are precisely articulated, falsifiable statements about the causal structure of the world. In political science, the greatest impact of recent computational advances has been to improve researchers' ability to test such theories. In a review of papers in the *American Journal of Political Science* (*AJPS*), the *American Political Science Review* (*APSR*), and the *Journal of Politics* (*JOP*) from 2018 to 2020, we identified 48 that employed statistical learning or other computational methods.[1] The vast majority of this work—over two-thirds—seeks to estimate a proxy for a concept in a causal theory that is not directly observable. Without this proxy, no empirical evaluation of the theory is possible.

While the use of proxies in social science is not new, our literature review highlights how computational methods have drastically increased their accessibility. For decades, the development of new proxies was a major effort, feasible only for well-funded research teams, that often attempted to make a new measure available to a broad community of scholars. For example, an enormous literature theorizes the effects of democratic institutions on a host of outcomes ranging from economic development to life expectancy. However, because "democracy" is not observable directly, any empirical test of these theoretical predictions must rely on a proxy. This necessity drove costly efforts utilizing large groups of expert coders, which have invited both widespread use and close scrutiny.[2] Similarly, to empirically test numerous theories about the origins and effects of legislator ideology, researchers commonly rely on a publicly available measure that was built from carefully derived statistical models based on application-specific functional-form assumptions about ideology and voting (i.e., NOMINATE; Poole & Rosenthal 1985). With the exception of multidimensional scaling methods for survey data and votes (Poole 2008), case-specific measurement models were, until recently, limited.

In a noteworthy paradigm shift, researchers now regularly estimate new proxies for individual studies, often from high-dimensional data for which traditional methods are inappropriate. At the

[1] **Supplemental Appendix Section A** describes our coding scheme, as well as the identified articles.
[2] For example, see Munck & Verkuilen (2002) for an evaluation of various measures of democracy, including Polity (Gurr 1974), Freedom House (2014), and others.

same time, implementing computationally intensive parametric models has become considerably easier with the advent of languages like Stan (Carpenter et al. 2017) and the vast computational power now available to researchers. Advances in statistical learning now allow researchers to flexibly estimate proxies without application-specific knowledge, using increasingly rich data sources and generic statistical models that adapt to the data at hand.

Despite these technological advances, there remain several fundamental research design considerations that receive little attention from researchers conducting analyses with learned proxies. Throughout this review, we use "proxy," as opposed to "measure," to emphasize that many such variables are substitutes that imperfectly estimate the underlying theoretical concept. At a high level, this slippage can stem from three sources: (*a*) measures often fail to fully capture all aspects of the underlying concept, (*b*) they often contain some level of purely random noise, and (*c*) they are often systematically contaminated by other factors besides the concept of interest. While an extensive literature on measurement has focused on improving validity by eliminating these sources of error, resource-constrained researchers often do not have the luxury of perfecting their proxy variables, particularly when the construction of these variables constitutes just one of many stages in the research process. Determining how to proceed in the face of these inevitable imperfections—the focus of this article—is therefore an important methodological question that confronts many applied researchers.

In the remainder of this review, we explain how these issues can bias both treatment effect estimates and standard errors. We then illustrate how scholars can use causal diagrams to reason about various sources of error and their implications in terms of statistical biases. It is well known that these diagrams constitute an easy-to-use tool for conveying the essence of social-scientific theories. What is less appreciated is that causal diagrams are also useful for concisely expressing the assumed quality of proxies used to approximate an underlying true concept, as well as for indicating potential sources of contamination. By writing down concrete assumptions in this easily digestible form, analysts can then apply well-established rules to determine which conclusions can be rigorously supported, while avoiding implausible parametric assumptions about functional form and the distribution of random errors. Without such parametric assumptions, which are generally difficult to defend, analysts generally cannot recover accurate quantitative estimates of theorized effect sizes. However, we show that in many common research settings, analysts *can* reliably evaluate the qualitative existence of these effects and determine their direction. That is, despite random measurement error and possible systematic error due to contamination by other factors, analysts can nonetheless rigorously assess whether treatment variables causally lead to the theorized increase or decrease in outcome variables. We provide numerous examples of research settings with proxied treatments, outcomes, and confounders in which such conclusions can be supported, along with straightforward procedures analysts can use when confronted with more complex scenarios.

2. INTEGRATING MACHINE LEARNING TECHNIQUES WITH SOCIAL SCIENTIFIC THEORY

Rigorous social scientific theories are statements about the causal structure of the world (Pearl & Mackenzie 2018). That is, they assert that a dependent variable, Y, would or would not have unfolded differently if an independent variable, X, had been hypothetically modified. Such theories are distinct from empirical predictions that X will be *associated* with Y, in that they posit an explanation for *why* empirical associations appear: for example, because X has a direct effect on Y, because it has an indirect effect through some intermediate factors, or because X and Y are both influenced by some common cause that produces a spurious correlation.

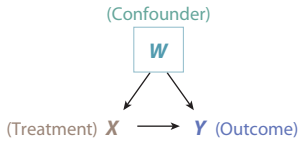

Figure 1

Theorized causal structure. The figure depicts a causal theory in which treatment X has an effect on outcome Y, but estimation is complicated by a common cause W that must be adjusted for (indicated with a rectangle) to recover $X \rightarrow Y$. Subsequent figures consider scenarios in which X, Y, or Z cannot be directly observed and must instead be noisily measured.

Well-articulated theories are collections of statements about (*a*) the set of factors that are theoretically meaningful and (*b*) how these factors might influence one another. These statements can be concisely expressed in the form of a causal diagram depicting each factor, with arrows representing influence relationships; a generic example is given in **Figure 1**. What causal diagrams do not convey is perhaps as important as what they do. Critically, causal diagrams do not make implausible claims about precisely how X affects Y. For example, they do not state that "the effect of increasing X by 1 unit is that Y will increase by an average of 2.5 units" or that "X_1 and X_2 have linear effects on Y and do not interact." In complex social scientific settings, analysts rarely have enough knowledge to theorize such rigid and specific functional forms. Instead, these relationships must be flexibly estimated from data.

Causal diagrams have proven invaluable to the social sciences, guiding both qualitative process tracing (Waldner 2015) and quantitative analyses (Keele et al. 2020) in the evaluation of social scientific theories. Classic references such as *Causality* (Pearl 2009) offer clear-cut guidelines for diagrammatically assessing alternative explanations that must be ruled out before analysts can draw firm conclusions. As a simple example, in the scenario of **Figure 1**, it can be seen that analysts must account for the common causes (or confounders, W) before estimating the theorized effect, $X \rightarrow Y$.[3] Without adjusting, analysts cannot rule out the possibility that observed associations between X and Y might be due to confounding, and the theorized causal effect might be nonexistent. In this review, we consider the complex issues that arise when analysts seek to evaluate their theories using indirect measures of key theoretical constructs—an increasingly common practice in social scientific research that uses rich, newly available data to proxy for concepts that were previously difficult to operationalize. We outline the types of conclusions that can and cannot be rigorously supported when analysts use learned proxies in common research settings, as well as a set of rules to help guide analysts confronted with more complex scenarios.

The fundamental problem that proxy-based research seeks to address is that theoretical concepts in the social sciences are often abstract and lack precision (Weber 2017). Consider the ideological bias, or slant, of media outlets. A staggering volume of research examines the origins of media bias, as well as its effects on subsequent social phenomena (Puglisi & Snyder 2015). At the time of writing, searching for "media bias" on Google Scholar yielded over 26,000 search results. Yet, media bias is not directly observable under any study design. Rather, this underlying true concept generates noisy and imperfect observed signals according to some generally unknown process. These imperfect signals may include (*a*) which politicians a given newspaper chooses to endorse (Ansolabehere et al. 2006), (*b*) the textual similarity between language used by media

[3]In potential outcomes notation, we have $X = X(W)$ and $Y = Y(W,X)$; the causal quantities of interest are taken to be various aggregations of or contrasts between the conditional average treatment effects, $\mathbb{E}[Y(x', w) - Y(x, w)|W = w]$.

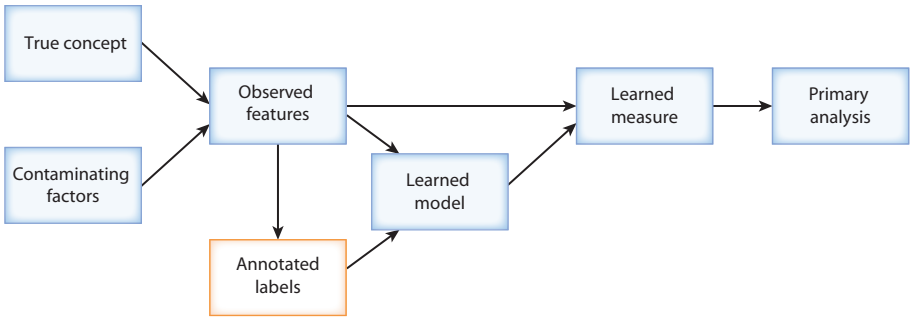

Figure 2

Overview of computational measurement. Each observation is associated with a specific value for the true concept of interest, which may be a confounder, a treatment, or an outcome. This attribute cannot in general be observed directly, but auxiliary information provides some signal about its value. However, these signals may be contaminated by additional factors; for example, if the attribute being measured is the treatment, the observed signals may contain not only treatment-related information but also contamination from the confounder. The attribute of interest may be annotated for a subset of units (indicated with gray text) based on observed signals. Annotations may contain errors or perfectly correspond to the true concept; if they contain errors, these errors may be independent or influenced by contaminating factors. After annotations are obtained (not obtained), supervised (unsupervised) machine learning models are trained—either on the observed signals directly or, more commonly, on a reduced representation that may result in the loss of information. The learned model is applied to observed signals for all units. The resulting estimates constitute the learned measure, which is then incorporated into a primary analysis.

outlets and members of Congress (Gentzkow & Shapiro 2010), or (*c*) the way an outlet covers certain issues (Larcinese et al. 2011). The absence of direct, high-quality data on the underlying concept greatly complicates the task of evaluating theories of media bias.

How do researchers use computational methods to address these challenges? **Figure 2** graphically depicts the typical workflow. As already noted, analysts have access to observed signals that convey noisy information about the concept via some process that is generally unknown. These signals potentially capture not only the true concept but also contaminating factors, discussed in Section 2.2. To map these signals back to the true concept of interest, researchers typically convert them into a reduced format that is amenable to analysis, then apply an assumed measurement model to obtain a predicted value of the true concept. It is critical to recognize that the model used for measurement is, at best, a simplified representation of the unknown process by which the true concept manifests in observed signals. Moreover, contamination of the signals used to proxy the true concept can lead to systematic errors that must be carefully considered when analysts draw conclusions. The construction and validation of measurement models, including with machine learning methods, has been the subject of much work (Adcock & Collier 2001, Grimmer & Stewart 2013). We briefly review this extensive literature before turning to the question of how measures should be used in subsequent, theoretically motivated analyses—a key component of the social-science workflow that has received far less attention.

2.1. Challenges with Computational Measures of Latent Variables

Measurement models are rich and varied, ranging from panels of human experts to keyword-based binary classification rules and trained neural networks. Here, we illustrate these and other choices confronting a researcher who is developing a computational proxy, using Martin & McCrain (2019) as a running example. Martin and McCrain study the effect of a sudden and widespread shift in television news ownership, which we denote as X, in which the conservative Sinclair

conglomerate acquired numerous media outlets in the United States. In this case, media consolidation was theorized to affect the unobserved concept of media slant, Y. Martin and McCrain use the measurement model of Gentzkow & Shapiro (2010), proxying media bias based on the similarity between (*a*) the text of each media outlet's news and (*b*) the text of partisan speeches in the Congressional Record.[4] We refer to the resulting predictions for each unit as the measure, $\hat{Y}$. We emphasize that the observed signal (which is often a rich information source, such as a television station's audiovisual stream) is conceptually distinct from the inputs to the measurement model (which can be lossy reductions, such as counts of various words obtained by an imperfect transcription).[5]

How might this proxy—textual similarity with the text of partisan speeches in the Congressional Record—differ from media slant, the unobserved latent concept of interest? For example, imagine that a researcher hopes to measure the slant of news articles covering *local* policy, which is generally not discussed in congressional speeches but still plausibly contains partisan bias. In this case, the similarity of these articles to the Congressional Record is not necessarily a good measurement model; its use requires researchers to assume that a model based on partisan speeches extrapolates well to topics not discussed in those speeches (local policy). If this assumption fails, relying on textual similarity between local news and congressional speech will yield unreliable results. Martin & McCrain (2019) address this concern by only applying the textual similarity measure to news segments covering national issues. However, if a researcher were interested in the ideological slant of local news coverage, they could alternatively rely on human annotators to inspect the observed signals (in this case, the text of news articles) and label the ideological slant of each document. This may be a difficult annotation task, because slant is often conceptualized as a continuous spectrum; in such cases, asking annotators to compare relative slant between document pairs, rather than report document-level slant values, can simplify the task substantially (Carlson & Montgomery 2017).

A benefit of using human annotators is that they often have tremendous contextual knowledge, understand ambiguous instructions, and can learn a large and flexible set of measurement models. Human annotators can also be given direct access to unstructured signals, such as audiovisual recordings of a television news broadcast, in their entirety. However, a limitation is that human annotators are expensive, so it may not be feasible to annotate every observation in the data. Moreover, even when annotations are available, humans are well known to exhibit prejudice and cognitive limits, meaning that annotated labels do not always reflect the underlying true concept. In some cases, key concepts may be difficult to precisely quantify even for experienced subject-matter experts, let alone the low-cost annotators that are often used for this task. Annotation errors, or differences between the truth and human labels, contribute to measurement error—a broader concept that refers to any difference between the true concept and a proxy (including machine predictions that may rely in part on human labels). Importantly, these errors exist at a conceptual level even when the underlying truth is unknown for all observations. As long as the true construct exists as theorized, then proxies must either deviate or not deviate from the underlying, unknown value, even though this deviation is not directly calculable. These deviations may either be purely random (e.g., accidental mislabeling) or systematic (e.g., higher slant scores for articles that are in ideological disagreement with the annotator). We return to this issue of

[4]This model can variously be thought of as a weighted, rescaled dictionary or as an instance of a supervised model trained on the Congressional Record and transferred to the domain of news.

[5]The practice of manually specifying informative inputs is referred to as feature engineering and can include stemming/lemmatizing of words, extraction of n-grams, and computation of interactions or other higher-order terms.

proxy quality later in this article, as it can introduce confounding and other statistical biases in subsequent analyses.

Setting aside challenges inherent in human coding, at a high level, supervised learning refers to the general approach of obtaining small to moderate amounts of annotation, then training a model that attempts to reconstruct the resulting labels based on some reduced feature set, such as word frequencies (Grimmer & Stewart 2013). The resulting learned model can then be cheaply applied to millions of unlabeled articles to obtain learned measures. Here, annotation errors can lead the model astray, but they are not the only problem; small training sample sizes or incomplete feature sets represent other sources of measurement error. However, annotation is not always needed, as indicated by the orange coloring of this step in the research workflow depicted by **Figure 2**. In contrast, unsupervised learning approaches attempt to identify latent clusters or dimensions that explain patterns in the observed signal without the need for human review. For instance, Poole & Rosenthal (1985) scale legislators according to voting patterns. Similarly, Slapin & Proksch (2008) scale documents according to word frequencies, based solely on co-occurrence patterns. These measurement models do not use human annotations to guide the process. Numerous variations (e.g., active learning, transfer learning) and hybrid approaches (e.g., semi-supervised learning, zero-shot learning) exist.

These computational methods represent powerful tools for mapping imperfect, messy, and high-dimensional signals about an unobserved theoretical concept to low-dimensional measures that can be used in statistical analyses. The trade-offs are well documented: Among other issues, such methods typically require moderate to large quantities of data to learn patterns without contextual knowledge; can overfit to limited data and memorize noise rather than learning generalizable patterns; and can learn only from the reduced space of features provided by analysts, which is typically more limited than the raw feature space available to human annotators. Various approaches have been developed to help address these obstacles to supervised learning, including cross-validation, transfer learning, and novel architectures that can ingest complex data. A full examination of these techniques is beyond the scope of this article. For a thorough review, including machine learning applications beyond those considered here, see Grimmer et al. (2021).

2.2. Using Computational Measures in Subsequent Analyses Will Bias Causal Estimates, But All Is Not Lost

Much of the prior literature on measurement focuses on improving the validity of the measure itself—that is, eliminating measurement error, especially from systematic sources (Adcock & Collier 2001). Yet, while measuring concepts has intrinsic value, we find that a far larger body of work is devoted to the next step of the scientific process: analyzing the origins and effects of the measured concept to improve our understanding of its broader social context. In our review of *APSR*, *AJPS*, and *JOP*, we found 48 papers that employed machine learning. Within this set, over two-thirds estimated a proxy for use in an empirical test of a causal theory. We found that this work on learned proxies fell into two categories. The first type (26 papers) made a primarily substantive contribution by developing a causal theory, then estimating a proxy variable in order to empirically test the theory. The second set (7 papers) made a primarily methodological contribution, focusing directly on the estimation and validation of a novel proxy variable for use in empirical tests of several causal theories.

These papers all confront a shared obstacle. How do we test theories that involve variables that we cannot directly observe? To do so, analysts employ a measurement model to create the learned proxy, which is then incorporated into a primary analysis. For example, Martin & McCrain (2019) conduct a regression of a proxy media bias outcome, $\hat{Y}$, on the theorized cause of media

consolidation, X, as well as other confounders, W. More generally, any theory that includes variables which are not directly observable is untestable without a learned proxy. In every other sense, proxy-dependent analyses are unremarkable, often employing common research designs intended to address classic threats to causal inference like unobserved confounding. For example, Martin & McCrain (2019) leverage a difference-in-differences design that compares acquired stations to other stations in the same market.

But while proxy-dependent designs often take seriously inferential threats like confounding, they commonly ignore challenges that are inherent in the use of a proxy variable. For example, random and systematic measurement error in a proxy can induce additional statistical biases in this primary analysis, with consequences that vary substantially depending on the quantity proxied and the precise nature of the error. But researchers often do not have the luxury of perfecting the measurement process due to constraints on time, personnel, and research funds. In many cases, noisy or contaminated observable signals can make it entirely impossible to obtain ideal measures of key theorized concepts. How can research proceed in the face of this challenge? We delineate how to reason about limitations of learned measures and how these limitations relate to the theorized causal structure. In Section 3, we then examine several common research settings and show that despite the statistical biases induced by imperfectly learned proxies, it nonetheless remains possible to draw meaningful conclusions about the theorized causal process.

We focus on lesser-known implications of measurement error and discuss what researchers can credibly conclude in the presence of bias resulting from this error. Specifically, we review how social scientists can draw conclusions about the existence and direction of a causal effect, even when the point estimate of that effect is almost certainly biased. Given that social scientists are generally most interested in demonstrating the existence of theorized effects rather than precisely quantifying the effect size, this result is encouraging. This goal is a particular form of causal discovery, a branch of causal inference that attempts to learn causal diagrams; in this context, researchers focus specifically on discovering a single theorized $X \rightarrow Y$ relationship, rather than considering all possible influence relationships. To the extent that a theorized effect is discovered, researchers may then seek to evaluate whether its direction, or sign, accords with theoretical expectations. (In Section 3, we formalize terminology for various senses in which effects can be described as positive or negative.) This research objective is distinct from the goal of causal estimation, which seeks to make precise quantitative statements about the magnitude of effects. Causal estimation appears in research on voter turnout and incumbency advantage, where the presence of an effect is already established with high confidence. As we discuss in this review, causal estimation is difficult in studies using learned proxies to approximate key unobserved steps in the theorized causal process.[6] In contrast, discovery and signing of an $X \rightarrow Y$ effect can be conducted under generally weaker assumptions about the types of contamination affecting a proxy.[7] **Figure 3*a*** depicts one possible causal structure, representing not only the theorized concepts but also a measurement process (in this case for the outcome Y). The $Y \rightarrow \hat{Y}$ arrow indicates a causal process by which the true outcome Y leads to the proxy $\hat{Y}$, compactly summarizing the entirety of the measurement workflow: the generation of observed signals; annotation of labeled units, if any; training of the model; and prediction of learned measures. It specifies this relationship without relying on untenable parametric assumptions. For example, the $Y \rightarrow \hat{Y}$ arrow does not state that measures are centered on the true concept, i.e., satisfy $\mathbb{E}[\hat{Y} - Y] = 0$. As

[6]Except in very specific cases that can be sensitive to violations of unverifiable assumptions about, for example, the functional form of the outcome.

[7]Causal discovery can be regarded as a precursor to causal estimation. Effects discovered with noisy proxies can highlight areas where improved measurement is necessary to obtain precise estimates.

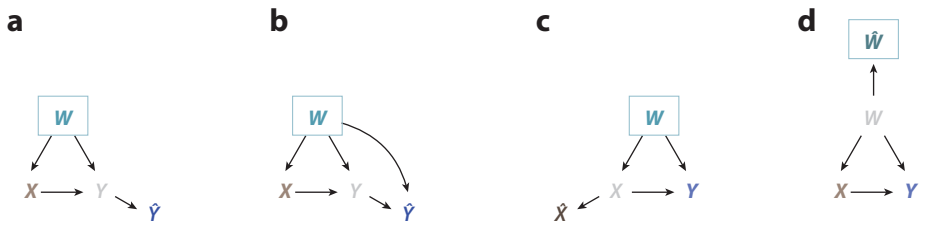

Figure 3

Causal structures of theory and measurement. Data environments correspond to the theory of **Figure 1**. Panels *a* and *b* illustrate settings in which analysts are unable to directly observe the outcome Y and thus must resort to a learned measure $\hat{Y}$ that is either (*a*) uncontaminated or (*b*) contaminated by confounders. The other two panels depict cases in which (*c*) the treatment X or (*d*) the confounders W cannot be observed, so that analysts can only adjust for learned proxies ($\hat{X}$ or $\hat{W}$).

the media slant illustration makes clear, such assumptions are often facially implausible. However, **Figure 3*a*** does encode structural assumptions in the absence of arrows from W or X to $\hat{Y}$, which state that the learned measure is uncontaminated—that is, free of influence from these factors, meaning that $\mathbb{E}[\hat{Y} - Y|W = w, X = x]$ is constant across all w and x.

This is a difficult requirement to satisfy; in many settings, analysts will be unable to defend the assumption that a proxy is uncontaminated. The main way to ensure it holds is to verify that the observed signal does not convey additional information about factors other than the true concept of interest. For example, in the media bias setting, contamination would occur if nonpartisan issues of rural interest (e.g., farm subsidies) tend to be discussed both in local news and by legislators representing rural districts. In this case, the confounder of rural-urban status would contaminate the media bias measure. In other words, rural-urban status (W) might distort the measure of media bias ($\hat{Y}$), above and beyond any influence that it might have on the true concept (Y). If this were true, **Figure 3*a*** would not be an accurate representation of the theory and measurement structure; **Figure 3*b***, in which W has a direct arrow to $\hat{Y}$, would be the correct representation. In other contexts, researchers may encounter scenarios where learned proxies must be used for the treatment of interest (X) or for key confounders (W); **Figures 3*c*** and **3*d*** depict these in turn.

When the observed signals are rich and unstructured (for example, when they contain text, audio, or images), it can be challenging to verify that they are free of contamination. It is therefore extraordinarily difficult to guarantee that unsupervised machine learning methods applied to such data sets will produce uncontaminated proxies of the true concept of interest. In the supervised setting, it is theoretically possible to obtain uncontaminated measures from contaminated features. When constructing a training data set, human annotators can be instructed to set aside their cognitive biases and label each unit according to objective scoring rubrics. For example, in the media bias case, annotators could be instructed that agriculture-related news should not be used as a signal of a news outlet's Republican leanings. But even if annotators perfectly adhere to these guidelines, a supervised measurement model that is regularized or incorrectly specified can often learn inappropriate shortcuts that reintroduce contamination, despite training on uncontaminated labels. We therefore recommend that analysts exercise caution. Measures should be thoroughly validated and probed for signs of contamination, e.g., by examining the predictive features used by the measurement model or by assessing whether agricultural keyword proportions continue to correlate with the media bias measure even after adjusting for obvious political keywords. However, definitive tests for contamination are often infeasible—in the above example, requiring countless model specifications and extensive keyword lists that range from "soybeans"

to "pesticide." We therefore recommend that when writing down assumptions in the form of a causal diagram, scholars should err on the conservative side by drawing arrows from all possible contaminating factors to the learned measure.

Having reviewed several challenges that arise when researchers use computational measures of a latent variable, we now explain precisely how researchers can make credible claims in the presence of bias resulting from these challenges. In the next section, we cover three cases in which computational measures are used to proxy the treatment, the outcome, or a confounder in a causal theory. We formalize and extend the common practice colloquially referred to as signing the bias, then use similar logic to show how valid inferences can be made about the presence of an effect even when point estimates are not point identified. By applying these rules, analysts can determine how various forms of contamination impact their ability to draw conclusions from available data.

3. ARTICULATING ASSUMPTIONS AND STRUCTURE

As noted in the previous section, researchers often use a measurement model to generate learned proxies when a true theoretical concept cannot be directly observed, as with media slant. In this section, we discuss the various research design complications that arise when researchers test causal theories with computational measures, depending on the theorized role of the proxied variable. We consider three cases in turn: where the learned measure proxies a treatment (Section 3.1), an outcome (Section 3.2), and a confounder (Section 3.3). We also consider certain settings where multiple factors are proxied simultaneously. We discuss prominent examples of each, drawing on recent research in international relations (Carroll & Kenkel 2019), comparative politics (Motolinia 2021), and American politics (Nyhan et al. 2012).

3.1. Learned Treatments

We first examine the case in which treatment, X, is approximated with a noisy learned proxy $\hat{X}$ (i.e., the error, $X - \hat{X}$, is nonzero). In this section, while discussing proxied treatments, we assume that the outcome (Y) and confounders (W, representing common causes of treatment and outcome) are perfectly observed. In discussing learned outcomes (Section 3.2) and confounders (Section 3.3), except where otherwise noted, we will consider cases in which only one variable is proxied and all other variables in the causal structure are observed without error. Finally, we assume that analysts use a model that does not make rigid functional form assumptions (or, less plausibly, that analysts know the exact functional form for the primary regression).

To illustrate the task of estimating causal effects of a treatment for which only a proxy is available, we point readers to a recent article by Carroll & Kenkel (2019), drawn from our review of machine learning applications. Carroll and Kenkel reexamine existing findings on the role of state power in conflict. As background, numerous theories predict that changes in a state's power will causally affect the chances of international conflict. In this study, the true treatment of interest, X, is state power—which is never directly observed. Carroll and Kenkel use machine learning to build a proxy measure of military power that improves on prior work. Drawing on data about the capabilities and outcomes of states involved in global military disputes, Carroll and Kenkel train a measurement model based on the material capabilities of the involved states and the outcomes of conflict, then demonstrate how their approach improves upon existing measures. Ultimately, the learned measure is used in the study's main objective: to revisit the findings of Reed et al. (2008) with these improved data. In this primary analysis, which uses a selection-on-observables design, Carroll & Kenkel (2019) find—contrary to the literature—that conflict is most likely when the state with the smallest benefits of war has a preponderance of power.

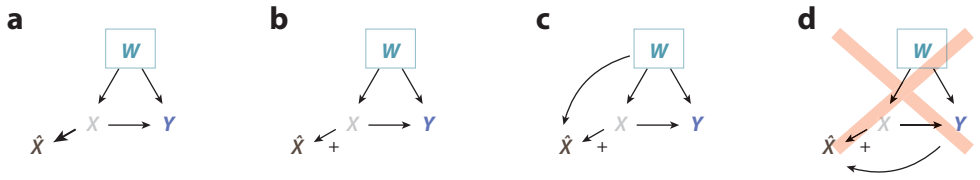

Figure 4

Learned treatments. In the settings shown, the true treatment X is unknown, but the proxy $\hat{X}$ can be estimated from auxiliary information. In all cases, both confounders W and outcome Y are known, and the analyst adjusts for W. (*a*) In a simple case where $\hat{X}$ is an uncontaminated proxy for X, a falsification test for the existence of an $X \rightarrow Y$ effect is possible. (*b*) In a similar setting with the additional (generally plausible) assumption that X has an on-average monotonic effect on $\hat{X}$, the sign of the $X \rightarrow Y$ effect can also be identified. (*c*) Even when $\hat{X}$ is contaminated by W, these results hold as long as W is adjusted for in a subsequent regression. (*d*) However, if the proxy is contaminated by the outcome itself, association between $\hat{X}$ and Y cannot be interpreted as evidence for the theorized $X \rightarrow Y$ effect.

With this example in mind, we now consider the general problem of drawing conclusions from proxied treatments and review related methodological work. First, it is well established that even when the measurement error $X - \hat{X}$ is independent noise, using a proxy $\hat{X}$ in place of X in a linear regression will result in attenuation bias (Wooldridge 2015, ch. 9.4). A number of theoretical results are available for linear and other parametric errors-in-variables models (we refer interested readers to Cheng & Van Ness 1999). And though scholars have known about bias induced by measurement error for decades, we find little mention of it in social science applications employing proxies. (In general, the use of imperfect proxies results in skewed estimates, with certain exceptions that we discuss below.) Moreover, further complications can arise if $\hat{X}$ is contaminated by additional factors, or if errors depend on the value of X itself, as commonly occurs.

When the true values of treatment X are available for a subset of the data (e.g., the proxy $\hat{X}$ is learned with a supervised model), Fong & Tyler (2018) offer a solution in contexts where the regression is known to follow a linear functional form. Intuitively, their approach uses $\hat{X}$ as an instrument for X. Specifically, in the first stage of regression, they relate X to $\hat{X}$ using the labeled data. In the second stage, where $\hat{X}$ is available but X is not, they use the full data to regress Y on $\hat{X}$. We discuss this procedure further in Section 3.3. However, a common concern is that linearity is unlikely to hold in complex social scientific settings.

We now review how analysts can still draw principled partial conclusions in the face of the above-mentioned issues, with weaker assumptions than those introduced by Fong & Tyler (2018). Specifically, after accounting for confounders W, analysts can conduct falsification tests—tests where the null hypothesis is the absence of an effect—about the causal influence of X on Y in **Figure 4*a–c***. Such tests are valid even in the presence of measurement error, because under the null hypothesis (i.e., in the absence of the $X \rightarrow Y$ arrow in the causal structure), there should be no association between $\hat{X}$ and Y (after adjusting for W).

Social scientists often seek to characterize the direction of causal effects, rather than simply testing null hypotheses about their nonexistence. By developing a statistical test with this objective in mind, we show how analysts can draw conclusions that would not otherwise be possible. To do so, we introduce the notion of signed causal diagrams (VanderWeele & Robins 2010), which build on the causal diagrams introduced in Section 2. In signed causal diagrams, researchers specify in their theory not only the presence of effects but also the direction of the effects. This practice is closely related to the common practice of signing the bias, in which researchers informally reason through how unobserved confounding might positively or negatively skew their results.

3.1.1. An introduction to signed causal diagrams. Researchers often informally state that one variable should have a positive or negative effect on another, but the meaning of these directions can be ambiguous. In this review, we focus on two possible assumptions about the direction of an effect, beginning with the assumption of average monotonicity. For two variables A and B, positive (negative) average monotonicity simply assumes that on average, as A increases, B either increases (decreases) or stays the same. Formally, the assumption of positive average monotonicity states that $\mathbb{E}[B(a') - B(a)] \geq 0$ for all $a' > a$. In a signed causal structure, we simply modify the arrows that we introduced in Section 2 to indicate whether the theorized effect is positive or negative. If A and B have a causal relationship satisfying positive average monotonicity, we label the corresponding arrow as $A \xrightarrow[+]{} B$, indicating that A has a nonnegative effect on the average value of B.[8] When discussing proxied outcomes later in this section, we show how a stronger condition, distributional monotonicity, is sometimes needed to justify conclusions about the direction of an effect. If positive distributional monotonicity holds, we write $A \xrightarrow[++]{} B$, indicating that increasing A will increase every quantile of B (including, e.g., the median of B). Formally, this is a statement about first-order stochastic dominance, requiring that $\Pr[B(a') \leq c] \leq \Pr[B(a) \leq c]$ for all $a' > a$ and all c. Some readers may be familiar with yet another type of signed effect, unit-level monotonicity, an even stronger assumption that we do not use in this review. This assumption states that if A is increased for any unit, B will also increase or stay the same for that unit.[9] These conditions are nested within one another: Unit-level monotonicity implies distributional monotonicity, which in turn implies on-average monotonicity. **Figure 4** presents several possible causal structures describing theory and measurement by means of signed diagrams indicating on-average monotonicity, the weakest and most plausible of the above monotonicity assumptions; we primarily focus on results involving this assumption.

3.1.2. Using signed causal diagrams for proxied treatments. Usefully, when learning $\hat{X}$ from data that are informative about X, analysts can generally assume that $X \rightarrow \hat{X}$ satisfies positive average monotonicity. This assumption requires that when X is larger, the estimated $\hat{X}$ will also tend to be larger, on average. In the context of learned proxies, we consider this to be a weak assumption; it is generally satisfied when well-calibrated machine learning models are used. This assumption can also be empirically assessed whenever the proxy is learned from labeled data. To do so, the researcher can simply train the model on a fraction of the labeled data (a training set) and inspect the accuracy of predictions in the remaining labeled data (a test set) by generating predicted values for the test set from the model learned in the training set. If average monotonicity holds, then predictions should correlate with the labeled values (which are known in the test set).

When average monotonicity between X and $\hat{X}$ is satisfied and W is correctly adjusted for, as in **Figure 4*b***, any positive association between $\hat{X}$ and Y implies a positive $X \rightarrow Y$ effect (VanderWeele & Hernán 2012). By the same logic, this is true for negative associations, which imply negative effects. Perhaps surprisingly, this also holds in the setting of **Figure 4*c***, in which the learned treatment $\hat{X}$ is contaminated by confounders W. At first glance, this contamination may appear to be problematic, as the measurement error will generally be associated with the outcome (as both $\hat{X}$ and Y are affected by confounders W). However, because analysts adjust for W, this concern is in fact unwarranted. Conditioning on W controls for the noncausal relationship

[8] If other parents of B exist, this must hold conditional on all possible values of these parents.

[9] Readers may be familiar with strong monotonicity from the "no defiers" assumption of Angrist et al. (1996) in the instrumental-variables setting. Formally, unit-level monotonicity $A \xrightarrow[+++]{} B$ states that $B_i(a') \geq B_i(a)$ for all $a' \geq a$ and all units i.

between $\hat{X}$ and Y that results from contamination from W. Specifically, controlling for W blocks two noncausal alternative explanations—termed "backdoor" paths by Pearl (1995)—from $\hat{X}$ to Y.[10] The first alternative explanation is that X does not have an effect on Y, but X is spuriously associated with Y due to confounding by W. Because $\hat{X}$ is influenced by X, this then also manifests in a spurious association between $\hat{X}$ and Y. This possibility, which is present in **Figure 4*a–c***, can be concisely expressed as $\hat{X} \leftarrow X \leftarrow W \rightarrow Y$. The second alternative explanation is that the measurement $\hat{X}$ is directly contaminated by the confounders W (i.e., that $\hat{X} - X$ is influenced by W), denoted $\hat{X} \leftarrow W \rightarrow Y$; this appears only in **Figure 4*c***. Both backdoor paths can be eliminated by adjusting for W, thereby breaking the chain of association (we refer interested readers to Pearl 2009 for a more comprehensive introduction to these concepts). We caution that if $\hat{X}$ is contaminated by Y itself, as in **Figure 4*d***, then association between the two clearly cannot be interpreted as evidence of a causal effect of X on Y.

However, the reverse is not true. Failure to find an association between $\hat{X}$ and Y does not necessarily indicate that no $X \rightarrow Y$ effect exists. In addition to standard issues of power in null hypothesis testing, there is the added issue that a poorly learned $\hat{X}$ may have no, or vanishingly little, association with X; this problem compounds with any power limitations that would arise in a non-proxied primary analysis. In other words, a lack of detectable association may be due to $X \rightarrow \hat{X}$ as well as $X \rightarrow Y$. We return to additional issues around uncertainty in Section 4.

3.2. Learned Outcomes

We next turn to the case when the true outcome Y is unobserved and analysts seek to draw causal inferences from a noisy learned proxy $\hat{Y}$. There are now countless examples of such applications, including every application of text analysis using topic proportions—an unsupervised measure based on observed term frequencies—as an outcome measure.[11] Several possible causal structures depicting theory and measurement are given in **Figure 5**.

Here, we highlight one prominent example to illustrate the concept. Motolinia (2021) studies the effect of allowing reelection on legislator provision of particularistic legislation. The theory states that for a legislator who is seeking votes and deciding where to allocate their effort, providing particularistic legislation will yield the most votes due to its targeted focus on constituent services. To identify the effect of reelection incentives, Motolinia uses a difference-in-difference design leveraging a staggered reform to elections in Mexico, which lifted a ban on reelection. Here, X is the ability of a politician to run for reelection, which is perfectly observed. In this case, confounders W are accounted for with state and month-year fixed effects. However, Y, the amount of particularistic legislation proposed, is not directly observed. To estimate the effect of the institutional transition, Motolinia must generate a measure $\hat{Y}$ of the outcome. To do so, Motolinia (2021) fits a correlated topic model (Blei et al. 2007) to legislative session transcripts, then classifies the resulting topics according to the legislation type. First, topics are grouped according to whether the legislation is procedural (e.g., protocol, voting rules), general (benefits all constituents), or particularistic (benefits a fraction of constituents). Motolinia validates this measure with extensive qualitative inspection and by confirming that the measure varies predictably in contexts where the

[10] This adjustment is straightforward when W is discrete (so that the association can be tested within levels of W) or when W's contribution to Y is additively separable from X's contribution [(i.e., when $E[Y|W, X] = f(W) + g(X)$)]; it may be difficult when W is continuous and interacts with X.

[11] Examples include approaches that explicitly couple the measurement and inferential processes, like the structural topic model (Roberts et al. 2013, 2014, 2016a), as well those that separately learn a topic model with, for example, latent Dirichlet allocation (Blei et al. 2003).

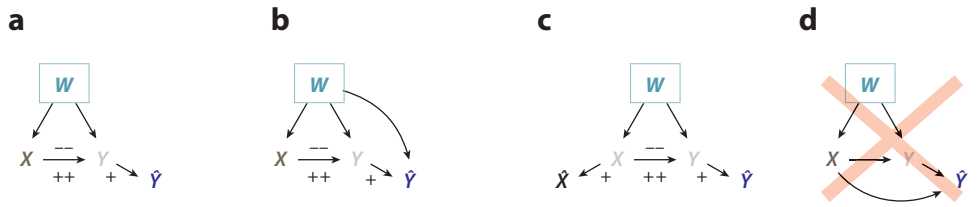

Figure 5

Learned outcomes. In the settings shown, the true outcome Y is unknown, but the proxy $\hat{Y}$ can be estimated from auxiliary information. In all cases, both confounders W and treatment X are known, and the analyst adjusts for W. (*a*) In a simple case, Y has an on-average monotonic effect on an uncontaminated proxy $\hat{Y}$, and the $X \rightarrow Y$ effect is known to exhibit distributional monotonicity. In this case, positive (negative) distributional monotonicity in $X \rightarrow Y$ is guaranteed to produce weakly positive (negative) $\text{Cov}(\hat{X}, \hat{Y}|W)$, and the sign of the $X \rightarrow Y$ effect can be identified. (*b*) This result holds even when $\hat{Y}$ is contaminated by W, so long as these contextual factors are adjusted for in a subsequent regression. (*c*) The same result still holds when X is also imperfectly, but on-average monotonically, learned. (*d*) However, if the proxy is contaminated by the treatment itself, association between X and $\hat{Y}$ cannot be interpreted as evidence for the theorized $X \rightarrow Y$ effect.

theory suggests it ought to (a test of face validity). The core outcome of interest Y is the proportion of particularistic legislation, and the proxy $\hat{Y}$ used in the regression is the estimated proportion according to this procedure.

We now demonstrate more generally how analysts estimating effects with a learned proxy, $\hat{Y}$, obtain the correct sign of the effect on the unobserved Y. It is well known that when the learned proxy is correct on average (i.e., when the error $\hat{Y} - Y$ has a conditional mean equal to zero), there is no bias in the point estimate (Wooldridge 2015, ch. 9.4). That is, using an unbiased $\hat{Y}$ in place of Y in a regression on W and X is equivalent to simply adding noise to the outcome. However, such perfect conditions rarely hold in practice; for example, when the proxied outcome is binary, the zero conditional mean assumption is violated if the learned model is more likely to misclassify a "true zero" as a "predicted one" than a "true one" as a "predicted zero" (i.e., if misclassification is asymmetrical). This is commonly the case, especially when one value of the outcome is more common than the other.[12] Moreover, if the measurement error $\hat{Y} - Y$ depends on treatment X, standard Gauss-Markov assumptions are violated and estimates will be biased.

We now describe conditions under which analysts can draw partial conclusions about the sign of $X \rightarrow Y$, even when Y is imperfectly observed, beginning with the structure of **Figure 5*a***. As before, we find the assumption of average monotonicity on $Y \underset{+}{\rightarrow} \hat{Y}$ to be generally plausible and empirically verifiable. Unfortunately, this assumption alone is not generally sufficient to assess whether $X \rightarrow Y$ is on-average positive or on-average negative. Because it is possible to construct examples where $X \underset{-}{\rightarrow} Y \underset{+}{\rightarrow} \hat{Y}$ and $X \underset{+}{\rightarrow} Y \underset{+}{\rightarrow} \hat{Y}$ both lead to positive correlations between X and $\hat{Y}$, analysts cannot conclude that X has an on-average positive effect on Y simply from observing that $\text{Cov}(X, \hat{Y}) > 0$. [For examples and detailed explanations, we direct interested readers to VanderWeele et al. (2008) and VanderWeele & Hernán (2012).]

Instead, a stronger condition, distributional monotonicity, is required. It has been shown that $X \underset{++}{\rightarrow} Y \underset{+}{\rightarrow} \hat{Y}$ always leads to a positive correlation between X and $\hat{Y}$, and similarly that $X \underset{-+}{\rightarrow} Y \underset{+}{\rightarrow} \hat{Y}$ always produces a negative correlation. Therefore, if the $X \rightarrow Y$ effect is known to exhibit distributional monotonicity, the sign of that effect can be inferred. We caution that when

[12] Even more implausibly, this perfect symmetry of misclassification must hold within all levels of X.

X and Y are continuous, distributional monotonicity in $X \underset{++}{\rightarrow} Y$ is a strong assumption that must be justified with domain expertise. However, an important special case is when both treatment and outcomes are binary, in which case average and distributional monotonicity are equivalent. In this case, this assumption is considerably simpler, and analysts can safely infer the sign of $X \rightarrow Y$ using the proxy $\hat{Y}$.

Next, we highlight two more complex cases in which analysts can nonetheless draw partial causal inferences from imperfect proxies. The first case is when W contaminates $\hat{Y}$, as in **Figure 5*b***; as discussed in Section 3.1, this is a minor issue because spurious association due to W can be adjusted for in the subsequent primary regression.[13] The second case is when learned versions of both $\hat{X}$ and $\hat{Y}$ are used in place of the true treatment and outcome, as in **Figure 5*c***. In this case, results generalize straightforwardly: Due to a technical result from VanderWeele et al. (2008), if $X \rightarrow Y$ is known to exhibit distributional monotonicity, then the conditional effect must share the sign of $\text{Cov}(\hat{X}, \hat{Y}|W)\ \text{Cov}(\hat{X}, \hat{Y}|W)$.[14]

3.3. Learned Confounders

Finally, we consider the difficult task of estimating causal effects by adjusting for an imperfectly learned confounder, $\hat{W}$, instead of the true concept, W. Again, illustrations of proxied confounders are plentiful in the social sciences. An especially prominent example is legislator ideal points (Poole & Rosenthal 1985), which researchers often wish to control for when explaining legislator behavior. Because ideology cannot be directly observed, political scientists construct proxies for ideal points with unsupervised scaling methods. There are at least two reasons for this. First, it would be very difficult to reliably hand-label each legislator on a continuous scale. Second, because all legislators routinely vote on the same bills, latent trait models are a reasonable way to project these votes down to one or two dimensions. Much research is devoted to the measurement of this variable, and we direct interested researchers to Clinton (2012) for further discussion.

Nyhan et al. (2012) demonstrate the importance of adjusting for ideological position when studying legislative voting. They examine the effect of controversial roll call votes, X—specifically, high-profile votes against the Republican-led healthcare reform in 2010—on subsequent electoral performance, Y. To estimate this effect, Nyhan et al. (2012) use a selection-on-observables design and note clear confounding by legislator ideal point W, which shapes both legislative positions and voter evaluation. After conditioning on estimated ideal points $\hat{W}$ and other confounders, Nyhan et al. (2012) estimate that votes against healthcare reform may have cost Democrats the majority in subsequent midterm elections.

We now consider the problem of drawing conclusions with proxied confounders. Intuitively, it is generally insufficient to simply treat $\hat{W}$ as if it were W, because the error $W - \hat{W}$ represents an unaddressed portion of the confounder that is associated with both treatment and outcome. Strategies nonetheless exist for recovering causal effects in certain settings, though we caution that available solutions are fragile in various ways described below. We further emphasize that common practice, which fails to account for the difference between $\hat{W}$ and W, deviates substantially from these solutions for estimating causal effects in the presence of proxied confounding.

[13] As noted in Section 3.1, the issue is fully resolved when W is discrete, or alternatively when $W \rightarrow Y$ and $X \rightarrow Y$ are additively separable.

[14] This is because the sign of the correlation induced by a path can be inferred by multiplying the signs of edges along that path when either (*a*) all edges display distributional monotonicity or (*b*) intermediate edges display distributional monotonicity and final edges are on-average monotonic.

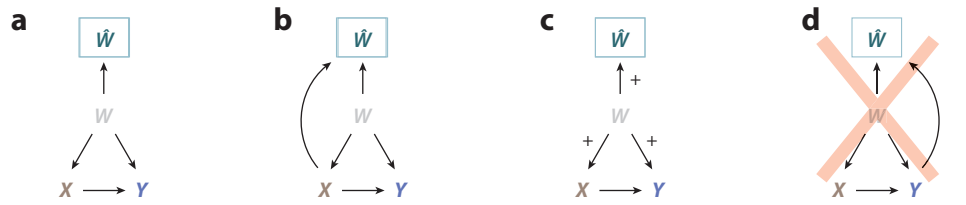

Figure 6

Learned confounders. In the settings shown, the true confounder W is unknown, but the proxy $\hat{W}$ can be estimated from auxiliary information. In all cases, both treatment X and outcome Y are known, but the analyst is only able to adjust for W. (*a*) A simple case in which $\hat{W}$ is an uncontaminated proxy for X. (*b*) A case in which $\hat{W}$ may be contaminated by X. Merely controlling for $\hat{W}$ is insufficient to obtain an unbiased estimate of $X \to Y$ in these cases. However, the methods described by Kuroki & Pearl (2014) and Miao et al. (2018) can in principle recover these effects if certain conditions are satisfied. (*c*) A case in which a true confounder W has a positive, on-average monotonic effect on both X and Y. In this case, an observed negative association between X and Y implies that a negative $X \to Y$ effect exists and is sufficiently strong to overpower the positive association induced by confounding. This remains true whether or not analysts adjust for $\hat{W}$. The reverse holds for negative confounding and positive association between X and Y. (*d*) A case in which the proxy is contaminated by the outcome itself, so causal effects are difficult to recover.

We begin by examining the simple setting of **Figure 6**. Greenland & Lash (2008) and Kuroki & Pearl (2014) establish that when W and $\hat{W}$ are discrete, causal effects are nonparametrically identified if analysts know the error mechanism, the distribution $p(\hat{w}|W = w)$—in other words, the pattern of correct and incorrect proxy values that arise from each possible true value. The basic idea is that when this error mechanism is known, it can be used in combination with the observed distribution of proxy values to back out the unobserved distribution of underlying true values.[15] This procedure can then be applied within each level of X and Y.

Approaches that rely on quantifying the error distribution are particularly attractive in supervised learning, where analysts following best practices already evaluate models in held-out validation sets. This validation set enables an unbiased estimate of the required information essentially for free, allowing analysts to recover the causal effects of interest. Indeed, in binary classification, widely used evaluation metrics—true and false positive rates—correspond exactly to $p(\hat{w}|W = w)$. Fong & Tyler (2018) build on this intuition in the linear case, developing a general method of moments estimator that simultaneously estimates the error distribution and the primary regression.

Results on proxied confounding also hold for the case when the learned confounder is contaminated by the treatment, as in **Figure 6*b***. However, if $\hat{W}$ is contaminated by Y, then the $X \to Y$ effect cannot be recovered. For this reason, Fong & Tyler (2018) recommend explicitly excluding Y from features used to train machine learning models. Even so, contamination can creep into the learned measure through numerous channels, including (*a*) learned models that inappropriately leverage correlates of Y; (*b*) model misspecification, as discussed in Section 2; or (*c*) contamination of observed signals that influence human annotations.

Kuroki & Pearl (2014) and Miao et al. (2018) extend these results to the challenging case when the true confounder is completely unobserved, as in unsupervised learning. A review of these techniques is beyond the scope of this article; for an overview of causal inference with proxy

[15] For analysts to do so, $\hat{W}$ must be sufficiently informative about W. The condition is violated if two confounder values, w and w', produce the same proxy distribution, $p(\hat{w}|W = w) = p(\hat{w}|W = w')$. This could occur if, for example, limited signals are incapable of distinguishing between two classes, or if there is a "ceiling effect" beyond which increasing W no longer affects $\hat{W}$.

confounders, see Tchetgen Tchetgen et al. (2020). However, we emphasize that these methods are substantially more involved than the common two-stage practice of fitting an unsupervised measurement model and then controlling for the result in the primary regression—a procedure that does not, in general, consistently recover causal estimates. Broadly speaking, consistent estimation of $X \rightarrow Y$ in the presence of imperfectly measured confounding is an extremely difficult task. Kuroki & Pearl (2014) note poor finite sample performance of these methods. Importantly, in the settings we examine, asymptotics depend critically on the size of the validation set; sampling errors in $\hat{p}(\hat{w}|W = w)$ can lead to substantial errors even when the primary analysis is based on infinite data. Tchetgen Tchetgen et al. (2020) also note issues with numerical instability, though these can be partially addressed with additional parametric modeling assumptions.

4. ACCURATELY REPORTING UNCERTAINTY

In the preceding section, we describe how imperfect proxies lead to biased point estimates. This raises an obvious question: Do imperfect proxies also lead to biased statements about uncertainty? In short, the answer is "yes." We now explain how analysts can draw principled conclusions despite these challenges.

In general, the common practice of using learned proxies as if they directly reflect the underlying true concept will bias standard errors downward, leading to overconfident conclusions that may fail to replicate. This is because standard procedures only account for uncertainty due to sampling variability in the primary analysis (the second stage of a proxy-based research workflow, which occurs after fitting the measurement model and estimating a proxy). What standard practice fails to account for is the fact that the learned measurement model (the first stage) is *also* estimated with a sample of data, introducing variability into the resulting proxy and therefore also contributing to overall uncertainty.

In other forms of research, such as analyses with missing data, it is well known that ignoring uncertainty from earlier stages (i.e., multiple imputation) leads to unreliable standard errors in subsequent regressions (Blackwell et al. 2017). Despite widespread awareness of this issue in related contexts, our review of published proxy-based work suggests that researchers rarely attempt to correct their standard errors. Among papers using computational methods in the *APSR*, *AJPS*, and *JOP*, the vast majority of papers analyzing learned proxies ignore the fact that these proxies are estimated with uncertainty (for details, see **Supplemental Appendix Section A**). The only exceptions were applications of the Structural Topic Model (STM) Roberts et al. 2013, 2014, 2016a). Interestingly, while substantive papers based on a proxy-dependent empirical test generally ignored uncertainty in the learned measure, methodological papers proposing a novel proxy often included a method for measuring uncertainty in the learned measure. For instance, Caughey et al. (2019) develop a proxy for mass policy ideology in Europe with a Bayesian dynamic group-level Item Response Theory model, from which uncertainty is easily extracted from the posterior estimates. But while this is common across Bayesian models, none of the identified papers incorporated this uncertainty into subsequent empirical analyses.

Before describing solutions to this issue, including the approach used by STM, we first review in more depth the sources of uncertainty that are often unaccounted for when learned proxies are used.

4.1. (Mostly) Ignored Sources of Uncertainty When Using Learned Proxies

Why do learned proxies lead to overconfident conclusions? Here, we briefly enumerate sources of uncertainty that, when ignored, lead to inappropriately small standard errors for the causal estimate of theoretical interest.

We begin by considering a supervised analysis in which the learned proxy is estimated from a training set, which is a sample from a population of possible training units. (Note that the same logic holds for unsupervised measurement models.) Our first source of uncertainty is the sampling variability that results from drawing one of many possible training sets. For simplicity of exposition, we assume that (*a*) annotators correctly label the underlying ground truth for each unit and (*b*) analysts recover a global maximum likelihood estimate for the measurement model, rather than a "local mode" that depends on randomly selected starting values (Roberts et al. 2016b). However, we note that in reality, these and other sources of nuisance variation can also undermine replicability of empirical conclusions.

Under these simplifying assumptions, given a particular training sample, applying a measurement model to this training set will lead deterministically to an estimate for the measurement model parameters. However, if a different training sample had been drawn, then the measurement model would have learned a different mapping from the observed signal to the concept of interest. This leads to a sampling distribution over learned measurement models.

These model parameters are in turn used to generate learned measures—whether $\hat{W}$, $\hat{X}$, or $\hat{Y}$—for each unlabeled unit in the primary analysis. Here, it is important to note that a slight change in the learned model (including changes due to a slightly different sample of training observations) will alter the generated proxy values for many units simultaneously. Put another way, over repeated sampling of the training set, learned measures in the primary analysis set are correlated across units. Moreover, units that have similar observed information will tend to shift similarly.

Our final source of uncertainty arises when learned measures are used in a primary analysis. The primary-analysis data set is also a sample from a broader population, producing another source of random variation. Apart from STM applications, every reviewed paper that reported a direct proxy-based test neglected training uncertainty; only uncertainty from the primary regression was reported.

Thus, a widespread methodological issue in existing work is the failure to adequately report uncertainty from the training process. There are numerous reasons why this issue has persisted. When analysts obtain pretrained machine learning models from third parties—e.g., commercial sources or other researchers—they may not know precisely how the sampling was done, and it may therefore be impossible to adequately account for uncertainty. For example, estimated sampling variances of model parameters might be reported, but not covariances. Similarly, if unit-level features are supplied to a cloud service, the service might respond with predictions and associated uncertainty for the unit-level learned measure, but cross-unit covariance is rarely reported by currently available services.

A simple *reductio ad absurdum* argument further illustrates the importance of training uncertainty for analyses based on $\hat{X}$ and $\hat{W}$ as well. If training uncertainty could in fact be safely ignored, in the limit, it would imply that binary classifiers could be trained on only two randomly sampled observations—one positive case and one negative case. The resulting model could then be used to learn measures for an infinite number of units. A primary analysis in this group would contribute no additional sampling uncertainty, due to its size. As a result, an analyst ignoring the training stage would claim perfect certainty in the results of their primary regression—an absurd claim, given that the entire analytic workflow hinges on a miniscule sample of two units. This illustration reveals that when properly accounted for, uncertainty vanishes only as both the measurement-model (first-stage) and primary-analysis (second-stage) data sets grow large. For this reason, we strongly discourage the widespread practice of ignoring training uncertainty (or, equivalently, reporting results "conditional on" pretrained models or learned measures based on their predictions). Because the causal theories being analyzed are ultimately about X, Y, and Z—not $\hat{X}$, $\hat{Y}$, and $\hat{Z}$, which

are merely proxies with no intrinsic causal role in the theory—analysts must take the underlying true concepts seriously.

As a final illustration, consider the use of a learned proxy, $\hat{Y}$. Correlation in $\hat{Y} - Y$ across units is functionally identical to correlation in the error term of a regression (as can occur in cluster randomized trials, for example, where units within a cluster may be simultaneously influenced by unobserved factors). It is easy to see that failure to account for correlated errors in the primary regression will typically lead to underestimates of uncertainty in the resulting estimates, much like failure to use clustered standard errors in a clustered design.

Given this challenge, how can researchers correct uncertainty estimates when using learned proxies? Next, we make a set of broad recommendations intended to aid analysts employing a range of designs.

4.2. Correcting Errors in Estimated Uncertainty

To represent uncertainty in the initial measurement stage, researchers can employ a range of common methods. Specifically, this uncertainty may be represented (*a*) with draws from a multivariate normal distribution, using point estimates and an estimated covariance matrix for the measurement-model parameters; (*b*) with draws from the joint posterior of parameters in a Bayesian analysis; or (*c*) with bootstrap draws of learned parameters, obtained by resampling of the training set and rerunning of the measurement model. Regardless of how it is obtained, each draw represents one possible measurement model that could have been learned; together, they approximate the spread of learned models that are plausible, given the finite training sample.

One improved and easy-to-implement method for reporting uncertainty follows the procedure of Treier & Jackman (2008). Take the first draw, $t = 1$, corresponding to one of the trained measurement models drawn as described above. Compute the proxy, e.g., $\hat{X}^{(t=1)}$, under this measurement model. Next, conduct the primary analysis using this proxy and extract the biased estimate of the quantity of interest, e.g., the $\hat{X}^{(t=1)}$ coefficient in a regression of Y on $\hat{X}$ and W. Uncertainty in this primary analysis can then be accounted for by taking P draws as above—i.e., by drawing from a multivariate normal approximation, drawing from a Bayesian posterior, or taking bootstrap draws. These draws approximate only the uncertainty in the primary analysis, taking the $t = 1$ proxy as given. The current standard practice stops at this point and, as a result, accounts for only primary-analysis uncertainty. In contrast, we recommend repeating the process T times, producing a total of $T \times P$ samples for the quantity of interest. Taking the 2.5th and 97.5th percentiles of the resulting distribution will produce an interval that reflects uncertainty from both measurement and primary analysis.

We caution that this procedure lacks many properties that analysts expect in traditional confidence intervals. In particular, due to the bias in point estimates that we discuss extensively above, it does not generally contain the true causal estimand in 95% of repeated samples. Despite this, it can be used in conjunction with the null-hypothesis-testing and effect-signing techniques developed above, while accounting for sampling variability in both stages of the analytic workflow.

The case of STM (Roberts et al. 2013, 2014, 2016a) illustrates an alternative, more complex approach for obtaining principled uncertainty estimates. Specifically, STM estimates a single model that encompasses both the initial measurement stage and the subsequent primary-analysis stage. This allows information to be passed back and forth between stages—e.g., the researcher can use patterns from the primary analysis to refine proxy predictions from the measurement model, and vice versa—leading to greater statistical power.

Relative to the sampling-based procedures describe above, the trade-off is that the joint modeling approach requires somewhat more technical familiarity and case-specific coding to implement.

However, joint modeling is increasingly feasible to implement in languages such as Stan. A variety of related methods for reporting uncertainty are also possible. For example, Knox & Lucas (2021) develop a hybrid two-stage approach for speech audio, in which coarse proxies are obtained from an initial training stage, then refined in an iterative process built into the subsequent primary analysis.

5. RECOMMENDATIONS AND CONCLUDING THOUGHTS FOR CREDIBLE ESTIMATES WITH LEARNED PROXIES

As our review demonstrates, it is now commonplace to generate learned proxies with computational methods as a first step in testing a causal theory. Though this approach has opened the door to numerous new and innovative studies, the use of imperfect proxies also presents challenges. To address these challenges, we now outline a series of best-practice recommendations for drawing principled conclusions from analyses using computational proxies of theorized concepts.

5.1. Explicitly State Your Causal Theory

As this review makes clear, formally specifying the theorized causal diagram has numerous benefits. Causal diagrams are concise and easy-to-use tools for communicating concepts to readers and clarifying the assumptions that underlie an analysis. Importantly, a well-specified causal diagram includes not only the theorized process but also a discussion of possible contamination sources and measurement-quality assumptions for proxies of unobserved variables. As we discuss above, clearly specified causal diagrams also help researchers reason about sources of error and assess what conclusions can be supported with a particular research design. Most notably, they reveal when null hypotheses and effect signs can be reliably tested.

5.2. Avoid Overclaiming Based on Biased Point Estimates

We show that primary analyses based on imperfectly learned proxies are almost always biased. Given this, researchers should be conservative in their interpretation by simply characterizing the sign of an effect, rather than making unsupportable claims about effect magnitude. If a researcher wishes to draw inferences about precise effect sizes, methods such as Duarte et al.'s (2021) offer a way to obtain bounds on possible effect sizes that account for the issues discussed here. Incorporating and expanding on this cautious approach to causal inference—whether by focusing on effect sign or through effect bounding—is an important avenue for future work by applied researchers and methodologists.

5.3. Test Your Assumptions

Researchers making assumptions (e.g., about the monotonicity of an effect) ought to assess their plausibility by drawing on past work, domain expertise, or empirical evaluation where possible. In particular, assumptions about on-average monotonicity in measurement, such as $X \underset{+}{\rightarrow} \hat{X}$, are straightforward to evaluate with procedures described in this review.

5.4. Always Assess and Report Measurement Performance

The performance of the measurement model undergirds any research design employing learned proxies. Proxies that are noisier or more skewed will tend to exacerbate the issues that we describe above. Among other issues, they can lead to false negative results: failure to find evidence in

support of a theory, even when that theory is true. In general, researchers should not trust results from any machine learning model until the performance of that model is suitably demonstrated. To demonstrate satisfactory performance, all applications should include a confusion table (i.e., cross-tabulation of true and predicted values) and other performance metrics obtained from a held-out validation set that was not used for training or parameter tuning. Finally, researchers should include measures of intercoder reliability, especially when annotating ambiguous labels. With limited resources, it may be inefficient to annotate each example in the labeled set repeatedly; instead, we encourage relabeling a sufficient number of cases to assess reliability. For example, with 2,000 training examples, it may be sufficient to hire a second coder to label only 100 for comparison.

5.5. Correct Your Standard Errors

As we note above, current practices for reporting uncertainty in proxy-based analyses are almost certainly biased, generally in a downward (anticonservative) direction. This is intuitive, given that analysts typically report uncertainty only for the second-stage model (the primary analysis, targeting the causal effect of interest) and neglect uncertainty and bias in the first stage (learning proxies). Unfortunately, without access to the learned model, it can be extraordinarily difficult to characterize how measurement error covaries between units. In the previous section, we describe methods for correcting this downward bias. Regardless which approach is used, analysts should seek to accurately report uncertainty from all stages of the model.

5.6. Compare Estimates in the Full Data to Estimates Using Only the Labeled Observations

We echo the observation by Fong & Tyler (2018) that, in the supervised case, a simple and consistent estimator exists: fitting a model using only the labeled data. This estimator is desirable for several reasons. First, and most obviously, it does not use proxies and therefore does not suffer from any of the sources of bias that we discuss in this article—at least, as long as the researcher can ensure that human labels reflect true, gold standard values for the underlying concept of interest. Second, substantial differences between the smaller-n unproxied analysis and the larger-n proxied analysis may indicate deeper issues that warrant further investigation. For instance, these estimates may diverge if the labeled data are not representative of the full data or if there are systematic errors in the classifier.

6. CONCLUSION

Our review highlights how advances in computational statistics are transforming research in the social sciences, primarily by allowing researchers to measure theorized concepts and use the resulting proxies in subsequent causal analyses. Yet despite the increasing prevalence of this research strategy, little methodological guidance is available for applied scholars. This is troubling because, as we note, the common practice of conflating proxies with the underlying true concept leads to biased point estimates and standard errors, undermining the conclusions drawn from this work.

Our analysis reveals that in spite of the recent computational revolution, core statistical obstacles faced by the discipline remain largely unchanged. In fact, our emphasis on precisely articulating theory and assumptions highlights that, ultimately, credible causal inference is about research design—same as it ever was. While new models may improve our ability to approximate previously unobservable concepts, no amount of computation can evaluate the plausibility of assumptions or prevent researchers from drawing unsupportable conclusions. It is thus unsurprising

that new research using new computational methods suffers from issues similar to those that ailed proxy-based studies decades ago.

But more optimistically, our review demonstrates how recent advances in causal inference can augment concurrent computational developments. By writing down their assessments of proxy contamination and measurement quality in the form of simple causal diagrams, analysts can now easily assess if a causal claim—whether about the existence, direction, or magnitude of an effect—is defensible. However, methodology in this area is far from complete. As computational social science continues to grow, much more work is needed to ensure that this rapidly expanding research area produces reliable scientific knowledge.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

For excellent research assistance, we thank Gechun Lin. For insightful comments, we thank Guilherme Duarte, Lucia Motolinia, Brandon Stewart, and Dustin Tingley.

LITERATURE CITED

Adcock R, Collier D. 2001. Measurement validity: a shared standard for qualitative and quantitative research. *Am. Political Sci. Rev.* 95:529–46

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444–55

Ansolabehere S, Lessem R, Snyder JM Jr. 2006. The orientation of newspaper endorsements in US elections, 1940–2002. *Q. J. Political Sci.* 1:393–404

Blackwell M, Honaker J, King G. 2017. A unified approach to measurement error and missing data: overview and applications. *Sociol. Methods Res.* 46:303–41

Blei DM, Lafferty JD. 2007. A correlated topic model of science. *Ann. Appl. Stat.* 1:17–35

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Carlson D, Montgomery JM. 2017. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *Am. Political Sci. Rev.* 111:835–43

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, et al. 2017. Stan: a probabilistic programming language. *J. Stat. Softw.* 76:1–32

Carroll RJ, Kenkel B. 2019. Prediction, proxies, and power. *Am. J. Political Sci.* 63:577–93

Caughey D, O'Grady T, Warshaw C. 2019. Policy ideology in European mass publics, 1981–2016. *Am. Political Sci. Rev.* 113:674–93

Cheng CL, Van Ness JW. 1999. *Statistical Regression with Measurement Error*. New York: Oxford Univ. Press

Clinton JD. 2012. Using roll call estimates to test models of politics. *Annu. Rev. Political Sci.* 15:79–99

Duarte G, Finkelstein N, Knox D, Mummolo J, Shpitser I. 2021. *An automated approach to causal inference in discrete settings*. Work. Pap., **https://arxiv.org/pdf/2109.1347.pdf**

Fong C, Tyler M. 2018. Machine learning predictions as regression covariates. *Political Anal.* 29:467–84

Freedom House. 2014. *Freedom in the World 2014: The Annual Survey of Political Rights and Civil Liberties*. Lanham, MD: Rowman & Littlefield

Gentzkow M, Shapiro JM. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78:35–71

Greenland S, Lash TL. 2008. Bias analysis. In *Modern Epidemiology*, ed. KJ Rothman, S Greenland, TL Lash, pp. 345–80. Philadelphia: Lippincott Williams & Wilkins

Grimmer J, Roberts ME, Stewart BM. 2021. Machine learning for social science: an agnostic approach. *Annu. Rev. Political Sci.* 24:395–419

Grimmer J, Stewart BM. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Anal.* 21:267–97

Gurr TR. 1974. Persistence and change in political systems, 1800–1971. *Am. Political Sci. Rev.* 68:1482–504

Keele L, Stevenson RT, Elwert F. 2020. The causal interpretation of estimated associations in regression models. *Political Sci. Res. Methods* 8:1–13

Knox D, Lucas C. 2021. A dynamic model of speech for the social sciences. *Am. Political Science Rev.* 115(2):649–66

Kuroki M, Pearl J. 2014. Measurement bias and effect restoration in causal inference. *Biometrika* 101:423–37

Larcinese V, Puglisi R, Snyder JM Jr. 2011. Partisan bias in economic news: evidence on the agenda-setting behavior of US newspapers. *J. Public Econ.* 95:1178–89

Martin GJ, McCrain J. 2019. Local news and national politics. *Am. Political Sci. Rev.* 113:372–84

Miao W, Geng Z, Tchetgen Tchetgen EJ. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* 105:987–93

Motolinia L. 2021. Electoral accountability and particularistic legislation: evidence from an electoral reform in Mexico. *Am. Political Sci. Rev.* 115:97–113

Munck GL, Verkuilen J. 2002. Conceptualizing and measuring democracy: evaluating alternative indices. *Comp. Political Stud.* 35:5–34

Nyhan B, McGhee E, Sides J, Masket S, Greene S. 2012. One vote out of step? The effects of salient roll call votes in the 2010 election. *Am. Politics Res.* 40:844–79

Pearl J. 1995. Causal diagrams for empirical research. *Biometrika* 82:669–88

Pearl J. 2009. *Causality*. Cambridge, UK: Cambridge Univ. Press

Pearl J, Mackenzie D. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books

Poole KT. 2008. The evolving influence of psychometrics in political science. In *The Oxford Handbook of Political Methodology*, ed. JM Box-Steffensmeier, HE Brady, D Collier, Vol. 10, pp. 199–213. Oxford, UK: Oxford Univ. Press

Poole KT, Rosenthal H. 1985. A spatial model for legislative roll call analysis. *Am. J. Political Sci.* 29:357–84

Puglisi R, Snyder JM. 2015. Empirical studies of media bias. In *Handbook of Media Economics*, Vol. 1, pp. 647–67. Amsterdam: Elsevier

Reed W, Clark DH, Nordstrom T, Hwang W. 2008. War, power, and bargaining. *J. Politics* 70:1203–16

Roberts ME, Stewart BM, Tingley D, Airoldi EM, et al. 2013. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, Vol. 4, pp. 1–20

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, et al. 2014. Structural topic models for open-ended survey responses. *Am. J. Political Sci.* 58:1064–82

Roberts ME, Stewart BM, Airoldi EM. 2016a. A model of text for experimentation in the social sciences. *J. Am. Stat. Assoc.* 111:988–1003

Roberts ME, Stewart B, Tingley D. 2016b. Navigating the local modes of big data: the case of topic models. In *Computational Social Sciences*, ed. RM Alvarez, pp. 51–97. New York: Cambridge Univ. Press

Slapin JB, Proksch SO. 2008. A scaling model for estimating time-series party positions from texts. *Am. J. Political Sci.* 52:705–22

Tchetgen Tchetgen EJ, Ying A, Cui Y, Shi X, Miao W. 2020. An introduction to proximal causal learning. arXiv:2009.10982 [stat.ME]

Treier S, Jackman S. 2008. Democracy as a latent variable. *Am. J. Political Sci.* 52:201–17

VanderWeele TJ, Hernán MA. 2012. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *Am. J. Epidemiol.* 175:1303–10

VanderWeele TJ, Hernán MA, Robins JM. 2008. Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology* 19:720–28

VanderWeele TJ, Robins JM. 2010. Signed directed acyclic graphs for causal inference. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 72:111–27

Waldner D. 2015. Process tracing and qualitative causal inference. *Secur. Stud.* 24:239–50

Weber M. 2017. *Methodology of Social Sciences*. New York: Routledge

Wooldridge JM. 2015. *Introductory Econometrics: A Modern Approach*. Mason, OH: Cengage Learning