CAUSAL GRAPHS IN POLITICAL METHODOLOGY

Doctoral thesis for obtaining the academic degree
Doctor of Social Sciences
(Dr.rer.soc.)

submitted by

SCHÜSSLER, JULIAN

at the

Faculty of Law, Economics, and Politics
Department of Politics and Public Administration

Konstanz 2020

Date of the oral examination: December 11, 2020

1st Reviewer: Prof. Dr. Peter Selb
2nd Reviewer: Prof. Dr. Anselm Hager
3rd Reviewer: Prof. Dr. Macartan Humphreys

19. Januar
Angst im Bureau abwechselnd mit Selbstbewußtsein. Sonst zuversichtlicher. Großer Widerwillen vor ›Verwandlung‹. Unlesbares Ende. Unvollkommen fast bis in den Grund. Es wäre viel besser geworden, wenn ich damals nicht durch die Geschäftsreise gestört worden wäre.

— Franz Kafka, Tagebücher.

This is not an exit.

— Bret Easton Ellis, American Psycho.

ACKNOWLEDGEMENTS

I would like to thank Peter Selb for his excellent supervision during my four years in Konstanz. He encouraged me to work on DAGs even though I had started the Ph.D. with a completely different topic in mind, and even though the subject is somewhat unusual for political science. It was him who eventually suggested working on it "officially". I had tremendous fun and benefited enormously from TAing for his lecture "Research Design I", and from having the opportunity to support him incorporating new causal graph material into it. The idea for parts of Paper 1 and for Paper 2 came from TAing for this lecture. Furthermore, Peter gave me the opportunity to teach my own seminar on Graphs. Finally, it was a great experience to have Peter as a co-author for this dissertation's last paper.

I would like to thank Anselm Hager and Macartan Humphreys for being part of my Ph.D. committee and providing me with valuable feedback. It is always fun and extremely insightful to discuss research ideas with Anselm. Macartan's commentary, coming from a prolific researcher in both substantive and methodological areas and an early and earnest adopter of graphs, is critical to me.

I would like to thank my co-authors Adam Glynn and Miguel Rueda for inviting me to a research stay at the Department for Quantitative Theory & Methods at Emory University. My stay, even though it was just six weeks long, was a phenomenally intense time. I learned a ton from both of them, and I am really proud of the product of our work (Paper 2 of this dissertation).

The Graduate School of Decision Sciences at the University of Konstanz allowed me to pursue this Ph.D. I would like to thank Jutta Obenland for being the true soul of this institution, being straightforward, and always looking out for people.

I would like to thank Michelle Jordan, Konstantin Käppner, Sascha Göbel, Nona Bledow, Sara Colella, Felix Gaisbauer, Philipp Kling, Hendrik Platte-Burghardt, Nico Gradwohl, Patrick Weber, Bihemo Kimasa, Dirk Streeb, Simone Stehle, Sandra Morgenstern, Jens Ihlow, Moritz Janas, Simon Roth, Julie Schnaitmann, Kevin Tiede, Mario Krauser, Phillip

Heiler, and Theresa Küntzler for being great colleauges at the GSDS.

Special thanks to Philipp Lutscher, Sebastian Hellmeyer, Max Heerman, Lukas Kawerau, and sorry for entering your office twice a day without knocking.

Very special thanks need to go out to Max Reinwald and Johannes Zaia and Max and Johannes together as the best office combo of all time. Together with Michelle, we made it to San Fransisco. That was crazy.

I would like to thank Konstantin Bätz for being a great flatmate and fitness trainer, and Enrico de Monte for being a great flatmate, if only for a short time. I would like to thank Maurizio for the sauna and clear conversations.

There were other people at the University who were important as friends and colleagues: Thomas Malang, Susi Breinlinger, Dirk Leuffen, Friederike Pförtner, Sarah Jabri, Steffi Klima, Maren Luy, Michael Herrmann, Michael Becher, Karsten Donnay, Tim Brackmann. Thank you.

I would like to thank the Studienstiftung des Deutschen Volkes for supporting my studies and my Ph.D. I would be in a very different spot if it wasn't for their support. I taught at the Studienstiftung's summer school in Leysin in 2018, together with Matthias Weierer and David Birke. Matthias and David, as well as all participants, were phenomenal. Thank you.

There are a few intellectual heroes of mine that have made this dissertation possible. I would like to thank Cosma Shalizi for leading me to graphs and hosting the world's best scientific website. I would like to thank Judea Pearl for being such an original and clear thinker. I would like to thank John Tsitsiklis for offering his life-changing course on Probability online. Moshe Hoffman on Twitter led me back to the social sciences, as strange as it sounds. Dax Werner was a constant companion either on Twitter or through his music.

I would like to thank my friends. Hendrik for the real, immediate conversations. Michael and Marlon for our therapeutic bike tours. David for so many interesting conversations. Alina for being Alina. Maya for being so amazing even though we know each other from school. Inga for that hike in Kyrgyzstan. Javier, Xiao Yu, and Chelsea for the connection across so much space.

Danke an Mama, Papa und Moritz.

Für Clara, mit der ich gerne Probleme löse.

ABSTRACT

Political scientists increasingly use causal graphs, specifically directed acyclic graphs (DAGs), to communicate identification assumptions for causal inference, but are reluctant to treat them as formal models. Their relationship to so-called "potential outcomes" has been largely unclear in both the applied as well as the methodological literature. This dissertation suggests that political scientists, as well as other empirical researchers, use causal graphs to communicate crucial assumptions, and in a second step to derive counterfactual and other independence assumptions from them.

In Chapter 2, I show that our understanding of existing analyses can be improved by using formal concepts from the causal graph literature. Specifically, I discuss how to systematically and transparently derive observable as well as a counterfactual assumptions from a given graph, and I apply these tools to four examples of published research. Here, I show how DAGs allow us to formally justify specification tests in causal mediation analysis, relax assumptions for complex observational studies as well as panel analysis, and illuminate the substantive content of assumptions in compliance modeling.

When using instrumental variables, researchers often assume that causal effects are only identified conditional on covariates. In Chapter 3—co-authored with Adam Glynn and Miguel Rueda—we show that the role of these covariates in applied research is often unclear, and that there exists confusion regarding their ability to mitigate violations of the exclusion restriction. We explain how existing adjustment strategies may lead to bias. We then discuss assumptions that are sufficient to identify various treatment effects, some of which are new, when the exclusion restriction only holds conditionally. In general, these assumptions are highly restrictive, albeit they sometimes are testable. We also show that other existing tests are generally misleading. Then, we introduce an alternative sensitivity analysis that uses information on variables influenced by the instrument to gauge the effect of potential violations of the exclusion restriction. We illustrate it by reanalyzing Spenkuch

and Tillmann (2017)'s analysis of Catholicism and voting in the Weimar Republic. Finally, we summarize our results in easy-to-understand guidelines.

In Chapter 4 Peter Selb and I demonstrate how DAGs can be used to encode and communicate theoretical assumptions about nonprobability samples and survey nonresponse, determine whether typical population parameters of interest to survey researchers can be identified from a sample, and support the choice of adjustment strategies. Following an introduction to basic concepts in graph and probability theory, we discuss sources of bias and assumptions for eliminating it in selection scenarios familiar from the missing data literature. We then introduce and analyze graphical representations of multiple selection stages in the data collection process, which highlights the strong assumptions implicit in using only design weights. Furthermore, we show that the common practice of evaluating adjustment variables based on correlations with sample selection or survey outcomes is ill-justified. Finally, we identify areas for future survey methodology research that can benefit from advances in causal graph theory.

The dissertation concludes with a discussion of these insights in relationship to parametric assumptions, robustness tests, political science theory, as well as the so-called "credibility revolution".

ZUSAMMENFASSUNG

Politikwissenschaftler verwenden zunehmend kausale Graphen, genauer gesagt gerichtete azyklische Graphen (Directed Acyclic Graphs, "DAGs"), um Identifizierungsannahmen für kausale Inferenzen zu kommunizieren. Sie zögern jedoch, sie als formale Modelle zu behandeln. Die Verbindung zum sogenannten "potential outcomes"-Ansatz is weitestgehend unklar, sowohl in der anwegeandten wie auch der methodologischen Literature. Diese Dissertation schlägt vor dass Politikwissenschaflter sowie andere empirische Forscher kausale Graphen benutzen um wichtige Annahmen zu kommunizierne, und in einem zweiten Schritt kontrafaktische und andere Unabhängigkeitsannahmen aus ihnen abzuleiten.

In Kapitel 2 zeige ich, dass unser Verständnis bestehender Analysen durch die Verwendung formaler Konzepte aus der Literatur zu kausalen Graphen verbessert werden kann. Insbesondere wird erläutert, wie beobachtbare sowie kontrafaktische Annahmen systematisch und transparent von einem gegebenen Graphen abgeleitet werden können. Anhand veröffentlichter Forschungsergebnisse wird dargestellt, wie dies konkret angewendet werden kann. Hier zeigt Kapitel 2, dass Graphen Spezifikationstests in der Kausalmediationsanalyse formal rechtfertigen und erklären können, Annahmen für komplexe Beobachtungsstudien sowie Panelanalysen zu lockern ermöglichen und substantielle Annahmen in der Compliance-Modellierung verdeutlichen.

Bei der Verwendung von Instrumentalvariablen gehen Forscher häufig davon aus, dass kausale Effekte nur nach statistischer Kontrolle von Kovariaten identifiziert sind. In Kapitel 3—verfasst mit Adam Glynn und Miguel Rueda—zeigen wir, dass die Rolle dieser Kovariaten in der angewandten Forschung oft unklar ist und dass Verwirrung hinsichtlich ihrer Fähigkeit besteht, Verstöße gegen die Annahme der Exklusion des Instruments zu beheben. Wir erklären wie bestehende Adjustierungsstrategien zu Verzerrungen führen können. Daraufhin diskutieren wir Annahmen, die ausreichen, um verschiedene Behandlungseffekte zu identifizieren, von denen einige neu sind, wenn

die Annahme der Exklusion des Instruments nur bedingt gilt. Im Allgemeinen sind diese Annahmen sehr restriktiv, obwohl sie manchmal empirisch testbar sind. Wir zeigen auch, dass andere bestehende Tests im Allgemeinen irreführend sind. Anschließend führen wir eine alternative Sensitivitätsanalyse ein, bei der Informationen zu vom Instrument beeinflussten Variablen verwendet werden. Diese erlaubt die Auswirkungen potenzieller Verstöße gegen die Annahme der Exklusion des Instruments zu messen. Abschließend fassen wir unsere Ergebnisse in leicht verständlichen Richtlinien zusammen.

In Kapitel 4 zeige ich gemeinsam mit Peter Selb, wie DAGs verwendet werden können, um theoretische Annahmen über nicht-Zufallsstichproben und Antwortausfall in Stichproben zu codieren und zu kommunizieren. Dies hilft zu bestimmen, ob typische Populationsparameter die für Umfrageforscher von Interesse sind aus einer Stichprobe identifiziert werden können, und Graphen unterstützen die Auswahl von Adjustierungsstrategien dazu. Nach einer Einführung in grundlegende Konzepte der Graphen- und Wahrscheinlichkeitstheorie diskutieren wir Verzerrungsquellen und Annahmen um diese zu eliminieren aus der Literatur zu fehlenden Daten. Anschließend führen wir grafische Darstellungen mehrerer Auswahlphasen im Umfrage-Prozess ein und analysieren sie. Dies unterstreicht die starken Annahmen, die bei der exklusiven Verwendung von "Design"-Gewichten auftreten. Darüber hinaus zeigen wir, dass die gängige Praxis der Bewertung von potentziellen Anpassungsvariablen auf der Grundlage von Korrelationen mit einem Antwortindikator oder den Umfrage-Variablen nicht gerechtfertigt ist. Schließlich identifizieren wir Bereiche für zukünftige Untersuchungen zur Erhebungsmethodik, die von Fortschritten in der Theorie der kausalen Graphen profitieren können.

Die Dissertation schließt mit meiner Diskussion dieser Einsichten in Bezug auf parametrische Annahmen, Robustheitstests, politikwissenschaftliche Theorie, sowie die sogenannte "credibility revolution".

CONTENTS

LIST OF FIGURES

Part I

INTRODUCTION

1

INTRODUCTION

Political science—and the social sciences in general—try to answer causal questions. Did the religious denomination of citizens in the Weimar Republic cause them to vote for the National Socialists (Spenkuch and Tillmann, 2017)? Are voters in the US demobilized politically even by short prison sentences, and does this depend on their race (White, 2019)? Does public support affect the stability of democracies around the world (Claassen, 2020)?

For such causal questions—all of which will be revisited in this dissertation—we have developed a more general understanding of "causal" methods based on non-experimental, observational data. Consequently, political science has shifted away from what has been called "traditional regression studies", which relied on "informally motivated sets of control variables", with little discussion of whether these are appropriate (Samii, 2016, pp. 941–942).

It is now clearer that for all kinds of inferences, causal or not, it is of central importance to have a clear grasp of the assumptions one invokes to gather evidence. And it may seem that a handful of methods and assumptions are well-understood and sufficient to establish causality (Angrist and Pischke, 2009). But very recently, political scientists, among others, have alerted us to the fact that things are generally more complex: it is difficult to experimentally identify causal mechanisms (Imai et al., 2011; Imai, Tingley, and Yamamoto, 2013), experiments may suffer from so-called post-treatment bias (Aronow and Miller, 2019; Coppock et al., 2019; Montgomery, Nyhan, and Torres, 2018), dynamic effects complicate panel analysis (Blackwell and Glynn, 2018; Imai and Kim, 2019), there are generalization problems in instrumental variable studies (Aronow and Carnegie, 2013), and administrative data may be of limited use due to the behavior of subjects under study (Knox, Lowe, and Mummolo, 2020). Survey researchers have pointed out that we need causal models to understand even simple descriptive inferences when data suffer from nonresponse (Groves, 2006). Taken together,

these developments suggest that there is no strict distinction between "causal" and "non-causal" methods and that for each approach, assumptions of potentially high complexity need to be invoked, sometimes more, sometimes less.

This dissertation argues that the way these complex assumptions are communicated in political science lacks clarity. By adopting a clearer approach, we will improve our understanding and our application of empirical methods. Specifically, the dissertation suggests to use the theory behind causal graphs (Pearl, 2009) as an overarching mathematical framework to communicate and work with causal assumptions. It applies this theory to important methods and problems, including causal mediation, instrumental variables, and nonresponse adjustment. It shows that the framework can consistently explicate the substantive content of assumptions, sometimes leading to broader applicability of methods, sometimes questioning their use fundamentally. The dissertation derives new statistical tests of assumptions while casting doubt on existing ones, and it develops a new sensitivity analysis for an underappreciated problem in instrumental variables estimation. In sum, this dissertation not only contributes to solving critical methodological problems in the social sciences but also suggests and applies a new framework to discuss and analyze these problems.

Originating in biology (Wright, 1920), causal graphs—and specifically directed acyclic graphs (DAGs)—are now a standard framework to think about causality in machine learning (Pearl, 2009; Peters, Janzing, and Schölkopf, 2017) and epidemiology (Greenland, Pearl, and Robins, 1999), and have recently also gained traction in quantitative sociology (Elwert and Winship, 2014a; Morgan and Winship, 2015), statistics (Maathuis and Colombo, 2015; Rothenhäusler, Ernest, Bühlmann, et al., 2018), and economic theory (Spiegler, 2016; Spiegler and Eliaz, Forthcoming). This dissertation specifically relies on the theory of structural causal models developed by Pearl (2009) that unifies the theory of causal graphs, structural equations, and potential outcomes.

How do political scientists traditionally communicate their assumptions, and what are the quantities they are interested in? Causal *quantities* are related to questions such as: What happens to turnout if we sentence every citizen to jail, as opposed to no citizen (the "average treatment effect", e.g., White, 2019)?

To what extent does a framing presented by a researcher affect policy preferences through individual evaluations of its importance, as opposed to evaluations of its content (indirect or mediation effects, Imai and Yamamoto, 2013)? A descriptive quantity answers questions such as: What is the association between socio-economic status and turnout (Lahtinen et al., 2019)?

Assumptions are typically formulated using terms like "exogeneity" or "ignorability". In fact, empirical research in many social sciences nowadays is judged to a large extent by whether such "exogeneity" conditions hold (Imbens, 2010). This culminates in articles like Schultz and Mankin (2019), which asks, "Is temperature exogenous?".

The latter article, as well as many others who use the term, does not state what exactly is meant by exogeneity. Nonetheless, the state of the art for defining quantities of interest and communicating such crucial assumptions in the social sciences, and specifically in political science, is to use "potential outcomes" or, equivalently, "counterfactual" variables. For example, in the study on religion and voting by Spenkuch and Tillmann (2017), which is reanalyzed in this dissertation (Chapter 3), the central assumption would need to be stated as "past observed religion is independent of the counterfactual vote for a fixed contemporary religion, conditional on other observed variables". This is not a very intuitive statement. For example, it sounds very similar to saying, "past observed religion is independent of the observed vote, conditional on observed contemporary religion and other observed variables". However, these two sentences are logically largely unrelated. Nonetheless, it is very common to depart from a counterfactual independence assumption and to appeal vague notions such as "as-if randomization" to justify it (Keele, 2015b). Can we improve this situation?

This dissertation uses causal graphs as a deeper, formal, yet intuitive framework to better understand and justify such assumptions. Consider the causal graph in Figure 1. Very briefly and applied to the situation in Spenkuch and Tillmann (2017), Z is past (17^{th} century) religion, D is contemporary (Weimar Republic) religion and Y is the vote share of the National Socialists (all on the county level). X are other measured variables, while U is not measured by the researcher. In Paper 2 of this dissertation, we show that the counterfactual independence assumption just described would hold if we took this graph as

a given assumption about the causal process behind the data. However, it is easy to show that the similar-sounding empirical independence statement would then generally be false.

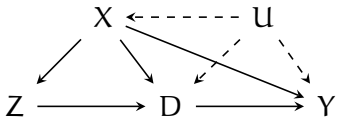

Figure 1: Graph analyzed in Paper 2. Conditional on X, Z is a valid instrument (is "exogenous") for the causal effect of D on Y.

On a fundamental level, this dissertation suggests that political scientists and other empirical researchers use causal graphs such as this one to communicate crucial assumptions. Only in a second step they should derive (not: assume) counterfactual and other independence assumptions. Using this approach, the dissertation shows that causal graphs can improve our understanding in almost every area of quantitative empirical inquiry: Causal mechanisms, complex observational studies, panel analysis, instrumental variable analysis, as well as survey research.

While they may strike one as intuitive and transparent, the use of causal graphs is somewhat controversial. A recent textbook by two eminent figures in causal inference states that they have found the approach not helpful and then chooses to ignore it completely (Imbens and Rubin, 2015, p. 22). At least until recently, there were also substantial disagreements between researchers using solely potential outcomes and those who also use graphs. Paper 2 of this dissertation shows—as others have shown for different scenarios, e.g.,Montgomery, Nyhan, and Torres (2018)—that controlling for an observed variable may introduce bias in specific (but relevant) situations, and should be avoided. Rubin, on the other hand, has written that "to avoid conditioning on some observed covariates...[is] nonscientific ad hocery" (Rubin, 2009, p. 1421). This dissertation hopefully contributes to political science and other disciplines leaving such needless schisms behind by showing that causal graphs and potential outcomes can—and should—be used together for solving important empirical problems.

In the next section, I summarize the contribution of the individual chapters of this dissertation. The section after that has a dual purpose. First, it introduces all mathematical concepts (beyond basic algebra), including the basics of DAGs, that are necessary to follow the arguments in this dissertation. This makes

it also possible for the reader to skip the repeated introduction of some of these concepts in the individual chapters.

Second, I discuss some conceptual underpinnings of this dissertation concerning estimation and identification, statistical and causal models, and how they relate to existing approaches.

The first subsection (1.2) starts with probability distributions and ends with a principled justification of linear regression as the conditional expectation function. Section 1.3 discusses the concepts of identification and estimation. The next section (1.4) contrasts the linear *regression* model with the the linear *causal* model, and introduces nonparametric causal models and their associated causal graphs. Next, I introduce the basics of directed acyclic graphs and d-separation (section 1.5).

Finally, I discuss briefly how causal inference is (not) presented and explained in two statistics textbooks by political scientists (Aronow and Miller, 2019; King, 1998) (section 1.6). This foreshadows some of the arguments in Paper 1. After Paper 2 and Paper 3, the final chapter (chapter 5) discusses some of the insights of this dissertation, especially with regards to the so-called "credibility revolution" (Angrist and Pischke, 2010).

1.1 SUMMARY OF CONTRIBUTION

1.1.1 *"Observable and counterfactual implications of causal graphs in political methodology"*

Paper 1, "Observable and counterfactual implications of causal graphs in political methodology", departs from the fact that while the potential outcomes approach has become the de-facto standard in political methodology since at least 2010, very recent methodological work also uses causal graphs, albeit informally. Furthermore, it is largely not clear whether and how the highly mathematical treatments using potential outcomes are connected to the vastly more intuitive causal graphs. To remedy this shortcoming, the paper explains how causal graphs can be treated as formal models, and how they can be used to derive both observable (testable) as well as counterfactual implications. It does so by making concrete progress in four methodological areas. First, it shows a statistical test for the central "causal independence of mediators" assumption in mediation analysis (Imai and Yamamoto, 2013) can be justified formally. Imai and

Yamamoto, as well as many other methodologists, claim that this assumption is not testable in principle. However, Imai and Yamamoto at the same time suggest a specific test of this assumption. I prove very generally, using graphs (both of which are also in Imai and Yamamoto (2013)), that this test is actually informative about deviations from the independence assumption, and discuss more specifically what a rejection may tell us.

Second, the paper shows graphical criteria can be used quickly and transparently to check whether a causal effect is identified. I analyze a graph from Samii, Paler, and Daly (2016) that depicts assumptions for a complex observational study involving multiple treatments and observed as well as unobserved confounders. Based on the concrete research context, I suggest that various of the assumptions depicted in this graph are incredibly strong, but can be relaxed without comprising identification, contrary to what the authors originally claimed.

Third, the article proves that central assumptions on the independence of unobserved confounders and observed covariates in panel settings (Imai and Kim, 2019) can be similarly relaxed. Again, the graph in question is taken from the original article. The basic intuition is similar to that one in the second application, although a formal derivation is considerably more complex, since standard graphical criteria such as the adjustment criterion (Shpitser, VanderWeele, and Robins, 2010) cannot be applied. And again, the original graph, when analyzed carefully, makes assumptions that seem too strong for most research applications: that all independent variables, including controls, are "exogenous", an assumption criticized by, e.g., Keele, Stevenson, and Elwert (2020).

Fourth, the article points out that the assumptions necessary for "compliance modeling" in instrumental-variables (IV) regression (Aronow and Carnegie, 2013) are, in the context of canonical IV graphs, equivalent to assuming no treatment-outcome confounding, which makes the use of IV superfluous. Here, neither the methodological nor the applied empirical literature uses causal graphs. Once one formalizes the problem using graphs, the strength of the assumption becomes apparent. Without further statistical development in this area, the use case for compliance modeling seems suspect. I discuss the application of compliance modeling on the effect of jail sentences on turnout by White (2019) to illustrate the problem.

Apart from the introduction of structural causal models (Pearl, 2009) to political science and the four individual methodological analyses, the article emphasizes throughout that causal graphs need to be analyzed carefully in any given substantive context, and that the important assumptions are the arrows that are left out.

1.1.2 *"Post-Instrument Bias"*

Paper 2, "Post-Instrument Bias" (co-authored with Adam Glynn and Miguel Rueda), shows that there are profound misunderstandings concerning the role of control variables in the applied (e.g., Kern and Hainmueller, 2009; Wucherpfennig, Hunziker, and Cederman, 2016) and methodological (e.g., Sovey and Green, 2011; Wooldridge, 2010) literature on instrumental-variables identification. The paper clarifies to what extent covariate control can aid in identification, and in more detail analyzes "post-instrument" variables: If we think that the instrument influences the outcome through another variable, and not only through the treatment, is it appropriate to simply control for this variable? Of all studies using instruments published in three top political science journals, we estimate that a quarter controls for such variables.

The paper shows that statistical control for such "post-instrument" variables may be necessary to identify a causal effect. Specifically, we discuss a new treatment parameter that can be identified in this context. But the assumptions, which we visualize using causal graphs, seem to be very restrictive, although we also prove that they are testable.

On the other hand, we show that that other existing tests—such as running analyses including or excluding the post-instrument variable—are misleading. Analysts need to commit to a causal graph a priori, and only in rare circumstances (that we explain) will these lead to a test.

Finally, it seems more likely that one cannot "control away" direct effects of the instrument effectively, and that the instrument is invalid. The paper, therefore, develops a semi-parametric sensitivity analysis for such situations that uses sample information to partially identify (bound) the causal effect of interest. The sensitivity analysis can also accommodate measurement error in the post-instrument variable. We illustrate it

by reanalyzing Spenkuch and Tillmann (2017). The application shows that it is important to allow for heterogeneity in causal effects, although the main inference remains robust.

1.1.3 *"Graphical Causal Models for Survey Inference"*

Paper 3, "Graphical Causal Models for Survey Inference" (co-authored with Peter Selb) shows that unit nonresponse, a quantitatively highly relevant departure from idealized random sampling, can be modelled using causal graphs. Even if the interest is in estimating means or associations (not necessarily causal effects), graphs can be used to understand biases that occur because some units systematically do not respond in surveys. Similarly, graphs can be used to derive and better understand assumptions for nonresponse adjustment such as "missing completely at random", "missing at random", and "missing not at random" (Rubin, 1976).

The paper then applies causal graphs to two problems in survey sampling and nonresponse adjustment. First, it analyzes multiple selection stages (e.g., sampling and response), and cautions against using only "design" weights in statistical analysis. Furthermore, it shows that the existing practice of estimating correlations of candidate adjustment variables with response indicators or survey variables (Kreuter et al., 2010; Peytchev, Presser, and Zhang, 2018; Sakshaug and Antoni, 2019) can be misleading. Valid adjustment variables may be uncorrelated with nonresponse indicators or the survey variable, because of offsetting paths. The paper concludes by discussing various areas where survey research may benefit from advances in the causal graphs literature.

1.1.4 *Summary*

In sum, these three papers advance our understanding of a wide variety of causal and statistical inference problems in political science and disciplines facing similar situations. All three papers explicitly or implicitly call for applied researchers and methodologists to embrace causal graphs as an essential tool to understand statistical analyses. Insofar as causal graphs are already in use—e.g., as in the recent papers by Imai and Kim (2019), Knox, Lowe, and Mummolo (2020), and Montgomery,

Nyhan, and Torres (2018)—the papers suggest to treat them as formal models, and to use rules and theorems based on graphs to justify statistical tests and counterfactual assumptions cleanly.

There is also a variety of more immediate and practical implications. Paper 1 suggests that the specification test for mediation models suggested but unnecessarily qualified by Imai and Yamamoto (2013) can be used widely (whenever two or more possible mediators are measured). It also calls on researchers to more carefully inspect their assumptions, even if (or especially) when they are represented in a graph. Finally, it questions the utility of compliance modeling in instrumental variable studies.

Paper 2 alerts researchers to an underappreciated problem that may fundamentally affect many instrumental variables studies. It shows that intuitive "robustness tests" may be uninformative in this context, but shows a concrete alternative test. Finally, the sensitivity analysis that we develop can be applied by any researcher who faces a potential post-instrument variable. It joins several other sensitivity tools developed for other problems that rely on similarly weak assumptions (Cinelli and Hazlett, 2020; Imai and Yamamoto, 2013).

Paper 3 implies that one should not use "design weights" alone, and that looking at correlations between adjustment variables and nonresponse indicators or survey variables—a central step in articles such as Kreuter et al. (2010) or Sakshaug and Antoni (2019)—is not informative.

1.2 BASICS STATISTICAL CONCEPTS

The following sections introduce basic mathematical concepts that are used throughout this thesis. I assume the reader is somewhat familiar with the basic notion of random variables and probability distributions; therefore, some of the discussion is dense.

I denote random variables with an upper-case letter like Y, sometimes with a subscript to emphasize the connection to the unit analysis, e.g.,Y_i. The probability distribution $P(Y)$ can be interpreted as the distribution of the random variable in a (finite or infinite) population. For example, it could describe the realized votes for candidates across electoral districts. Then, e.g., $P(Y = \text{Red}) = 0.2$ means that 20% of all votes were cast for the

Red candidate or party. Here, Red is a specific realization of Y, which is generically denoted by a lower-case y.

Alternatively, Y could be the outcome of a stochastic process, e.g.,an estimator in a sampling process. For example, we could imagine drawing an infinity of finite samples from a population of units and computing a function of the realized values of the variables in that sample. The function itself is called an estimator. The value of such a function is usually called a statistic or an estimate. The resulting distribution of the estimator is often called the sampling distribution.

This may lead to some confusion as one usually deals with two distributions. First, there is the true distribution of the variable, which a priori has no connection to the research process. For example, we would like to describe the distribution of votes across electoral districts, but may only do so with some error if we only have a sample available. Second, if we repeatedly describe such distributions using a sample of data, our actual descriptions (estimates) have a sampling distribution, which we also may not be able to characterize exactly.

To make matters more concrete, we may be interested in the mean, expectation, or expected value $E[Y]$ of a variable Y. For discrete Y, this is defined as

$$E[Y] = \sum_y y \cdot P(Y = y), \tag{1}$$

where $\sum_y$ indicates summing over all possible outcomes y of Y. For continuous Y, the expected value is

$$E[Y] = \int_y f(y)dy, \tag{2}$$

where $f(y)$ is the probability density function of Y. Continuous variables play only a minor role in chapter 2 of this dissertation, and I will concentrate (without much loss of generality) on the discrete case.

Expectation is a linear operator, that is, for constants a, b, c and random variables X, Y, we have

$$E[a + bX + cY] = a + bE[X] + cE[Y]. \tag{3}$$

The joint probability function $P(Y, X)$ describes the probability that two (or more) random variables X and Y take on a specific outcome. It is the basis for concepts such as the covariance and correlation of two variables and like these, it is fundamentally symmetric: For all x, y, $P(Y = y, X = x) = P(X = x, Y = y)$.

The conditional probability $P(Y|X)$ describes how the probability of Y taking on a specific value depends on the value of X. There are at least two intuitive interpretations of this:

1. We imagine to subset the population or data along units with the same X, and then look at $P(Y)$ within this subset, or
2. We consider what our belief about Y—this is the subjective or Bayesian interpretation of probabilities—is after learning the value of X.

The latter interpretation is especially helpful for getting an intuition on statistical independence, defined as $P(Y|X) = P(Y)$ ("the random variable Y is independent of the random variable X"). Consider your belief about the age distribution in a given country, $P(Y)$. Now consider learning the income X of a person. If this does not change your assessment of that persons's age Y, for any income and any age, then we would say that the two variables are independent. The shorthand notation for statistical independence used in this dissertation (due to Dawid, 1979) is $Y \perp\!\!\!\perp X$.

A further basic and very useful rule of probability is the law of total probability. It reads

$$P(Y) = \sum_x P(Y|X = x)P(X = x). \tag{4}$$

That is, we can calculate the marginal distribution $P(Y)$ from the conditional (subgroup) distributions $P(Y|X = x)$, weighting each by the subgroup size $P(X = x)$ ("divide and conquer").

It is easy to generalize the concepts of a joint distribution and independence to the conditional case. That is, $P(Y, X|Z)$ is the joint distribution of Y, X conditional on the variable(s) Z. Y and X are said to be independent conditional on Z if $P(Y|X, Z) = P(Y|Z)$. The shorthand notation for this is $Y \perp\!\!\!\perp X|Z$.

There are a number of useful properties of conditional independence, some of which will be elegantly substituted by graph

theoretical concepts later (see Paper 1 for a more extensive discussion). Among these are Dawid, 1979; Pearl, 2009,

- Symmetry: If $X \perp\!\!\!\perp Y|Z$, then $Y \perp\!\!\!\perp X|Z$.
- Contraction: $X_i \perp\!\!\!\perp Y_i|Z_i$ and $X_i \perp\!\!\!\perp W_i|Z_i, Y_i$ imply $X_i \perp\!\!\!\perp Y_i, W_i|Z_i$.
- Independence of function of random variables: If $X \perp\!\!\!\perp Y|Z$ and U is a function of X, then 1) $U \perp\!\!\!\perp Y|Z$ and 2) $X \perp\!\!\!\perp Y|Z, U$.

Contraction is interesting in that it allows "chaining" independencies: If X is not informative about Y, and also not informative about W once you know (condition on) Y, then it is not informative about W generally (unconditionally).

The conditional version of the law of total probability reads

$$P(Y|X) = \sum_z P(Y|X, Z = z)P(Z = z|X). \tag{5}$$

That is if we restrict the attention a priori to a certain subset of the population (by conditioning on X), then the divide and conquer strategy using Z also happens completely conditional on X.

Most of statistics is interested in means (averages), correlations, and variances. I have already defined the unconditional mean (expectation) $E[Y]$. It is then easy to define the covariance of two variables as

$$cov(X, Y) = E[XY] - E[X]E[Y], \tag{6}$$

and the variance of a single variable as a special case thereof,

$$var(X) = cov(X, X) = E[X^2] - E[X]^2. \tag{7}$$

These two equalities play a central role in the derivation of the semi-parametric sensitivity analysis in paper 2.

The square root of the variance is called standard deviation, denoted σ_x. The correlation of two variables is defined as

$$cor(X, Y) = \frac{cov(X, Y)}{\sigma_x \sigma_Y}. \tag{8}$$

A further central concept in statistics is the conditional expectation

$$E[Y|X = x] = \sum_y y \cdot P(Y|X = x). \tag{9}$$

and the conditional expectation function

$$E[Y|X] = \sum_y y \cdot P(Y|X) = f(X). \tag{10}$$

The conditional expectation is a number, while the conditional expectation function is a function (of X).

The equivalent of the law of total probability for expectations is the law of iterated expectations. It states that

$$E[Y] = \sum_x E[Y|X = x]P(X = x) = E[E[Y|X]]. \tag{11}$$

I now use these concepts to introduce and justify linear regression. If X is discrete with outcomes $0, \ldots, x$, then one can always write

$$E[Y|X] = \alpha + \beta_1 I(X = 1) + \ldots + \beta_x I(X = x). \tag{12}$$

Here, $I(\cdot)$ is the indicator function that has value 1 if the condition in brackets is true, and 0 otherwise.

This equation merits some further analysis. First, note that the right-hand side is a linear function of the parameters α, β. One may, therefore, call this equation a linear regression (of Y on X), although there are also other definitions of regression (see below). Second, note that the equation is linear by virtue of the discreteness of the independent variable. The nature of the dependent variable Y does not matter at all. No further assumptions have been invoked. Third, the parameters are clearly defined as certain conditional expectations (or their differences). For example, it is easy to see that $\alpha = E[Y|X = 0]$ and $\beta_1 = E[Y|X = 1] - E[Y|X = 0]$, and so on. Fourth, all of the parameters are population quantities, defined independently from any estimation routine.

Fifth, we may bring the equation into a more familiar form by adding the random variable Y and rearranging, yielding

$$Y = \alpha + \beta_1 I(X = 1) + ... + \beta_x I(X = x) + (Y - E[Y|X]. \quad (13)$$

Renaming $\epsilon = Y - E[Y|X]$ yields

$$Y = \alpha + \beta_1 I(X = 1) + ... + \beta_x I(X = x) + \epsilon. \quad (14)$$

In the special case of a binary X, the correct equation (by definition, invoking no additional assumptions) is

$$Y = \alpha + \beta X + \epsilon. \quad (15)$$

This looks like a familiar linear regression. The regression error ϵ has a number of useful properties by construction. First,

$$E[\epsilon|X] = E[Y - E[Y|X]|X] = E[Y|X] - E[E[Y|X]|X] = E[Y|X] - E[Y|X] = 0. \quad (16)$$

The first equality follows from the definition of ϵ, the second from linearity of expectations, and the third equality from the definition of conditional expectation.

Second,

$$E[\epsilon] = E[Y - E[Y|X]] = E[Y] - E[E[Y|X]] = E[Y] - E[Y] = 0. \quad (17)$$

The first equality follows from the definition, the second from linearity of expectations, and the third from the law of iterated expectations. Together, this implies

$$E[\epsilon|X] = E[\epsilon] = 0. \quad (18)$$

The first equality specifically implies that X and ϵ are *mean independent*. This further implies that

$$cov(X, \epsilon) = E[X\epsilon] - E[X]E[\epsilon] = E[E[X\epsilon|X]] - E[X] \cdot 0 = E[XE[\epsilon|X] = E[X \cdot 0] = 0. \quad (19)$$

Here, the equality $E[X\epsilon] = E[E[X\epsilon|X]]$ follows from the law of iterated expectations. In the inner expectation, X is a constant

conditional on X, which is why we can use linearity of expectations and equate this with $E[XE[\epsilon|X]$.

This derivation of mean independence of X (independent variables) and ϵ (regression error) followed from the discrete nature of X, which further implied that $E[Y|X]$ is linear in the parameters.

More generally, whenever we equate $E[Y|X] = X\beta$ a priori, or specify any other (potentially non-linear) function for the conditional mean, such as $E[Y|X] = \frac{1}{1+exp(-X\beta)}$ for the logistic regression, the resulting regression error will be mean independent of X. This follows from defining $\epsilon = Y - E[Y|X]$ and the derivations in equations 16 and 17. Of course, the conditional mean function may be misspecified. In the case of discrete regressors, as shown above, a linear specification is always correct.

1.3 IDENTIFICATION AND ESTIMATION

The preceding discussion was on the level of random variables as population quantities. It did not discuss issues that arise when we only have a sample of data. When we stay on this population level and ask whether we can determine a certain quantity of interest, we are asking whether this quantity is *identified*.

An example from paper 3 may be helpful. In this paper, we ask under what circumstances we can learn about the distribution of variables in the whole population when these variable can be measured only in a segment of this population. The examples we discuss come from the area of survey research, where nowadays usually only 5 to 70% of the population of interest is willing to participate. We define an indicator variable S that is 1 if a unit would participate in a given survey. $P(S)$ is the distribution of this variable in the population. $P(Y|S = 1)$ is the distribution of a variable of interest Y—say, political preferences—among people that would participate in a survey. But often, the interest is in the overall distribution $P(Y)$, and these distributions may differ, for reasons that are explained in paper 3.

Accordingly, even if we were able to ask every person in the population to participate in the survey, we would only be able

to measure $P(Y|S = 1)$. In this sense, $P(Y)$ is not identified. One *identification assumption* to achieve this is to say that

$$Y \perp\!\!\!\perp S. \tag{20}$$

This is equivalent to saying that $P(Y|S = 1) = P(Y)$, and so we have immediately achieved identification. The left-hand side is what we assume to know, and the right-hand side is what we wish to know.

This may strike one as cheating: We first defined the problem as having information only conditional on $S = 1$, and then assumed that this simply does not matter. However, such independence assumption are at the heart of most identification strategies, and in this regard, this particular assumption has no special status. What is more relevant, and this is one of the main themes of this dissertation, is to be very clear about the substantive meaning of these assumptions, given a research context. This is one of the motivations to use causal graphs.

In sum, this dissertation uses a semi-formal notion of identification as being able to equate a quantity of interest (e.g., $P(Y)$) "uniquely" with a quantity assumed to be known a priori (e.g., $P(Y|S = 1)$). This is consistent with the definition used by Pearl (2009, p. 77).

This approach to identification differs from the very loose, but popular definition in the seminal textbook by Angrist and Pischke (2009). They define "an identification strategy to describe the manner in which a researcher uses observational data (i.e., data not generated by a randomized trial) to approximate a real experiment" (Angrist and Pischke, 2009, p. 7). This would mean that the argument above on finding $P(Y)$ is not identification, as no randomized trial is involved. More generally, this definition excludes many other interesting questions, such as generalizing experimental findings, where the challenge is to use a mixture of experimental and observational data to approximate *another* experiment (Pearl and Bareinboim, 2014).

The task of identification is logically distinct from the task of estimation. Estimability is usually conceived of as the existence of a consistent estimator for a quantity, that is, an estimator that converges to the true value of the quantity as the number of observations increases.

A quantity may be identified, but not estimable. Even seemingly simple quantities such as conditional expectation func-

tion cannot be estimated without further assumptions; one generally needs to assume "smoothness" (e.g.,Wasserman, 2013, p. 107, Peters, Janzing, and Schölkopf, 2017, p. 103; for an in-depth treatment, see Maclaren and Nicholson, 2019).

For linear regression, a consistent estimator exists under weak assumptions, e.g.,that observations are not "too dependent" and that covariances are finite (the linear function already is "smooth"). For more general problems, consistent estimators exist under similar independence assumptions and assuming the regression functions are continuous and have continuous derivatives (e.g., Shalizi, 2019, pp. 33, 43, 84–89, 491.)

Problems with highly dependent observations have arguably been somewhat neglected in causal inference. In the social sciences, such dependencies are likely to occur when the system under study has network properties. Lee and Ogburn (2020) have recently restated prominently that that such dependencies can lead to spurious associations in any given sample (although not over repeated samples).[1] Similarly, problems with possibly highly non-linear regression functions are not the main focus in applied causal inference, where linear regression dominates. An important exception are regression discontinuity designs, where the standard has long been to use nonparametric regression (Hahn, Todd, and Klaauw, 2001).

While identified quantities may not necessarily be estimable, the "reverse" case seems to be more likely: That we can estimate a regression consistently, but it is not equal to a (causal) quantity of interest. Finally, the case "not identified, but estimable" cannot exist. If a quantity is not identified, the problem is that is cannot be equated to a statistical quantity, and so we cannot even ask whether the latter is estimable, because it is not defined.

1.4 REGRESSION, STRUCTURAL MODELS, AND CAUSAL GRAPHS

So far, very little has been said on causation. Although the model

$$Y = \alpha + \beta X + \epsilon \tag{21}$$

1 Ogburn, VanderWeele, et al. (2014) and Aronow, Samii, et al. (2017) are recent causal inference perspectives on networks.

may suggest an asymmetric causation flowing from X to Y—from the "independent" to the "dependent" variable—the way the model has been defined does not rely on (nor necessarily communicates) causation. If this is not clear yet, one can evaluate the quantity $cov(X, Y)$ using this model:

$$cov(X, Y) = cov(X, \alpha + \beta X + \epsilon) = \beta var(X), \tag{22}$$

which yields $\beta = \frac{cov(X,Y)}{var(X)}$. The crucial step here is using the equality $cov(X, \epsilon) = 0$, which, as shown above, is a consequence of specifying the conditional mean function correctly, and always holds if regressors are discrete.

One might as well say that β is defined as this scaled covariance, and this indeed occurs in an even more general case that treats linear regression as the result of an optimization problem (Angrist and Pischke, 2009, p. 35). Since correlation (here: covariance) is not causation, the linear model above is not (inherently) causal. On the other hand, it is not useless: It provides a parsimonious description of how X relates to the mean Y.

So when is such a model causal? Pearl, based on others (refs), suggests one answer: *If we say so*. Specifically, he states that (Pearl, 2009, p. 135)

> the conditions that make the equation $Y = \beta X + \epsilon$ structural is precisely the claim that the causal connection between X and Y is β and nothing about the statistical relationship between x and ϵ can ever change this interpretation of β. Amazingly, this basic understanding...has all but disappeared from the literature, leaving modern econometricians and social scientists in a quandary over β.

I will later illustrate how this "quandary" plays out in an influential political methodology textbook.

For now, I note two things. First, in this quote, and in the rest of this dissertation, the word *structural* is synonymous with *causal*.

Second, it is perhaps unsurprising that there has been some confusion over the interpretation of β, because it really can and has been used to denote two wildly different things—the conditional expectation gradient as well as the causal effect of X on Y.

Some of the newer literature in machine learning (Peters, Janzing, and Schölkopf, 2017), therefore, writes the causal model explicitly as

$$Y := \alpha + \beta X + \epsilon. \tag{23}$$

Here, the assignment operator := is meant to imply asymmetric causality flowing from X and ϵ to Y, so that β is *defined* as a causal effect.

The modern literature, and most of this dissertation, is preoccupied with the general *nonparametric* causal model (here again using a standard equality sign)

$$Y = f_Y(X, \epsilon). \tag{24}$$

This equation is fairly general. It merely communicates that Y *could* be influenced by X, and also influenced by other unobserved variables ϵ. The interpretation of ϵ is "all other variables that influence Y when X is fixed". Note specifically that ϵ might be a vector of multiple variables; in the linear model above, ϵ is just a scalar (for each observation).

An interesting case is the model

$$Y_i = \alpha + \beta_i X_i + \epsilon_i. \tag{25}$$

If X_i is binary, this model is actually completely nonparametric.[2] All possible individual-level heterogeneity could be soaked up in either ϵ_i or β_i. If X_i is not binary, or if one introduces additional variables to the model, then the model may not be completely general. E.g., for continuous X_i, the model would imply that the effect of X is the same no matter which changes in X one looks at, for any given individual. However, allowing for arbitrary heterogeneity in causal effects across individuals seems generally attractive. Paper 2 uses such models to develop a sensitivity analysis, and calls the model "semiparametric". This is because there is some functional form assumption involved (fixing the individual, the equation is linear), but, on the other hand, no distributional assumptions on

2 Hahn, Todd, and Klaauw (2001), a foundational paper for regression discontinuity designs, switches between this and alternative but equivalent notations very transparently.

β_i or ϵ_i are invoked. Powell (1994) discusses the notions of parametric, semi-parametric, and nonparametric models, and shows that the distinction is not always clear. In this dissertation, following Pearl (2009), a nonparametric model is one involving no functional form or distributional assumption.

Nonparametric models such as eq. 24 have been investigated in econometrics for some time (Imbens and Newey, 2009). The structural causal models approach by Pearl, 2009 suggests that a useful abstraction of such models is a causal graph. In the case of the model in equation 24, we would need to complement it with a model for X. A simple option is to say that

$$X = f_X(\eta). \tag{26}$$

Here, we make an assumption: Y does not influence X, as it does not appear in the causal function.

The associated graph then is in Figure 2.

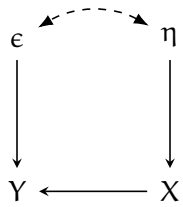

Figure 2: Causal graph associated with the causal models in equations 24 and 26.

It is hopefully intuitive how the translations between the equations and the graph happened: Whenever a variable occurs in the structural function for another variable, a directed arrow is drawn from the former to the latter. The bidirected arrow $\leftrightarrow$ between η and ϵ indicates that these errors may be correlated. They would be uncorrelated if were willing to assume that none of the unobserved causes of Y that are in ϵ are also unobserved causes of X, and vice versa. I will return to this assumption, which is critical. Having motivated causal graphs coming from regression and structural models, the next section takes them as given and introduces some basic graph-theoretic concepts.

1.5 DIRECTED ACYCLIC GRAPHS

The three graphs in Figure 3 are the fundamental directed acyclic graphs (DAGs). Any larger DAG can be constructed using these three graphs, plus the elemental direct path $X \to Y$. Once one understands how correlation and causation plays out in these three graphs, one can analyze (in principle) any larger DAG.

DAGs consist of variables, which are linked by edges (arrows). A *path* is a sequence of arcs that links one node to another, regardless of the direction of arrows. Retracing arcs or going through the same node twice is not allowed. A *directed* or *causal path* is traced out along arrows tail-to-head. If there is a directed path from one node to another, the former is said to be an *ancestor* of the latter, the latter a *descendant* of the former. A directed acyclic graph contains only directed arrows and no feedback loops (i.e., no variable is its own ancestor or descendant).

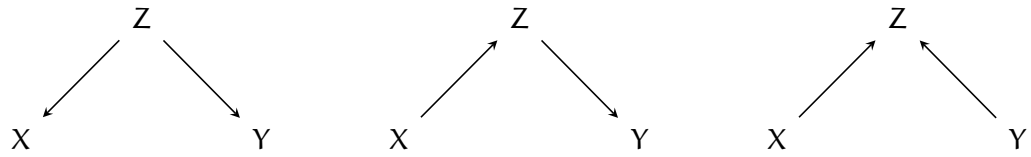

Figure 3: Basic causal patterns and d-separation: Z is a *confounder* (left), a *mediator* (center), or a *collider* (right) on the path between X and Y. In the latter scenario, the path is naturally blocked by Z, which entails that there is no statistical association between X and Y. Otherwise, paths are open and induce statistical association between X and Y.

In the left panel of Figure 3, Z is a common cause, or *confounder*, of X and Y. In this scenario, Z induces a statistical association between X and Y, although X and Y do not cause each other, which is indicated by the absence of an arrow between them. In the intermediate graph, Z is a *mediator*. This also has the consequence of producing a correlation between X and Y because the former influences the latter. In short, we say that these two paths are *open*. Finally, in the right panel, Z is a common effect, or *collider*, on the path between X and Y. In contrast to the two other cases, this does not produce a correlation between X and Y. Here, we say that the path is *blocked* by the variable Z.

These patterns of association reverse once we look at the distribution of X and Y conditional on Z, i.e., $P(Y|X,Z)$. In the left graph, Z is the only common cause of X and Y. Accordingly, for units with the same value of Z, the value of X is not informative about Y, and vice versa. In the intermediate graph, conditioning on Z also blocks the information flow from X to Y. X and Y are said to be d-separated conditional on Z. However, in the collider graph on the right, an association emerges. To understand why an example is helpful.

Consider two independent binary variables X and Y and a random variable Z that is the sum of X and Y. Therefore, Z can take on the values $\{0,1,2\}$. X and Y may be random coin flips, so knowing the value of X does not help in predicting Y. However, conditioning on Z means that we are told its value. Knowing that Z is 1, for example, and that X is 0, we know that Y has to be 1. Conditional on Z, X and Y are dependent or d-connected. Put differently, knowing the result of a process and the value of one of its independent inputs also lets us predict the value of the other input. The same mechanics apply if we happen to know the realization of a descendant of Z. For example, let D be a variable that takes on the value 1 when Z equals 1, otherwise 0 (so that it is a binary proxy for Z). Knowing that D equals 1 and that X equals 0 also leads to the prediction that Y equals 1.

In sum, in Figure 3, conditioning on the intermediate variable Z blocks the path between X and Y in the first two graphs but opens the path in the right graph. For deriving whether variables X and Y are (conditionally) independent in more complex DAGs, it turns out that one can just enumerate all paths between these variables. If all of these paths are blocked, perhaps conditional on other variables Z, then we say that X and Y are d-separated (conditional on Z) (Geiger, Verma, and Pearl, 1990).

If all of the variables involved are measured, this statement is testable: If Z d-separates X and Y, then $P(Y|X,Z) = P(Y|Z)$. For instance, if one commits to a specific regression model, the test involves regressing Y on X and Z; the coefficient on X should be zero (One could also use X as the dependent and Y as the independent variable). The only reason that the coefficients can be different from zero is that the DAG is incorrectly specified and there is at least one open path.

1.6 CAUSAL INFERENCE IN TWO POLITICAL METHODOLOGY TEXTBOOKS

The confusion between regression and causal models that I discussed in Section 1.4 also appears in political methodology. For example, an influential textbook (King, 1998, p. 8) starts with the formulation

$$Y_i = x_i\beta + \epsilon_i. \tag{27}$$

It then goes on to say

> [W]e generally assume that x_i and ϵ_i are independent. We have become used to thinking in these terms, but conceptualizing ϵ_i and what it correlates with is not as easy to explain in terms close to one's data and theory.

The point is then not further developed. Specifically, no explanation is given what ϵ_i or β actually refer to. The reader is left to wonder whether this model is causal or a non-causal. This is unfortunate, but not uncommon for textbooks (Chen and Pearl, 2013).

The discussion resurfaces once more in a later footnote (King, 1998, 166 fn.3), which defines

$$\epsilon \equiv Y - \mu, \tag{28}$$

where μ generally signifies the unconditional or conditional mean of Y.

As discussed previously, defining the error as deviations from a conditional mean function implies that the fundamental object of interest is this function, the regression of Y on X.[3] Because of this, it seems that the King (1998) textbook is fundamentally interested in association, not causation, and that it has no formal apparatus to differentiate between the two concepts.

While many, perhaps most articles published today in the social science are still in the tradition of King (1998), the formal discussion of causality gained prominence starting in 1995 with

3 The further discussion in King (1998) (for example, the foundational equation 1.2 on p. 8) is very much consistent with this interpretation.

a series of papers on the interpretation of instrumental variables in what is called the potential outcomes framework or Rubin Causal Model (Angrist and Imbens, 1995; Angrist, Imbens, and Rubin, 1996). This swapped over to research in political methodology a bit later and is mirrored, for example, in work by Kosuke Imai and co-authors (Imai, Keele, and Yamamoto, 2010, e.g.). A recent textbook in this tradition is Aronow and Miller (2019) to which I now turn. Here, the explicit focus on causal inference is clear.

Formally, the innovation is to differentiate observed variables Y_i and X_i from unobserved potential outcomes $Y_i(x)$. Usually, this is motivated with a binary variable X_i indicating some kind of medication. Then, the potential outcome $Y_i(1)$ indicates an individual's health when she takes the medication, and $Y_i(0)$ indicates her health when she does not. We conceive of the difference between these potential outcomes as the causal effect of the medication on this person's health. The "fundamental problem of causal inference" (Holland, 1986) is that we observe only one of these outcomes. Without further assumptions, we can therefore not infer any kind of causal effect.

An obvious advantage of this framework is that it is much clearer on the *definition* of causality and causal effects. Can the same be said about the assumptions that are invoked to learn about causal effects? The textbook by King (1998) alluded to the assumption that X_i and ϵ_i be independent, although it also said that was unclear what ϵ_i actually referred to.

In the potential outcomes tradition, central assumptions are also independence assumptions that involve potential outcomes. The most common one is sometimes called "strong ignorability" or often simply "(the) conditional independence" assumption:

$$Y_i(x) \perp\!\!\!\perp X_i | Z_i. \tag{29}$$

How is this assumption explained? The textbook by Aronow and Miller gives two explanations (Aronow and Miller, 2019, p. 248). First:

> The conditional independence assumption states that, among units with the same measured characteristics, the types that receive the treatment and the

> types that do not are exactly the same in terms of their distribution of potential outcomes.

Second:

> In other words, the conditional independence assumption implies that, after accounting for observable background characteristics, knowing whether or not a unit received treatment provides no additional information about a unit's potential outcomes.

It is fair to say that these explanations are very similar to each other, and that both are merely close restatements of the mathematical statement $Y_i(x) \perp\!\!\!\perp X_i | Z_i$ in English. At the start of this introduction, I mentioned a variant of such a statement, applied to the situation in Spenkuch and Tillmann (2017). It seems implausible—and in fact, it very rarely happens—that applied researchers actually communicate substantive arguments using this language.

The next chapter of this dissertation shows how to bring together DAGs and potential outcomes in the context of published methodological and substantive research, and how it can facilitate communication and analysis. Inter alia, it also shows that the error term King (1998) talks about and the potential outcomes that Aronow and Miller (2019) discuss are very closely related. This suggests that the doubt in King (1998) about the interpretability of the error term directly applies to potential outcomes as well. The chapter therefore suggests to replace conditional independence assumptions as primitives with causal graphs that are then used to derive such independence assumptions transparently. This is also, very broadly, what Papers 2 and 3 suggest, via concrete applications in instrumental variables identification and nonresponse adjustment.

Part II

RESEARCH ARTICLES

2

OBSERVABLE AND COUNTERFACTUAL IMPLICATIONS OF CAUSAL GRAPHS IN POLITICAL METHODOLOGY

Much of quantitative political science has been transformed by the adoption of the counterfactual approach to causality. In this approach, quantities of interests—causal effects—as well as central assumptions are formulated using counterfactual or "potential", as opposed to observed outcomes (Keele, 2015a; Samii, 2016). Groundbreaking contributions from political methodology to this interdisciplinary endeavor cover topics such as the analysis of causal mechanisms (Imai et al., 2011), causal inference in networks (Aronow, Samii, et al., 2017), and conjoint analysis (Hainmueller, Hopkins, and Yamamoto, 2014).[1]

Even more recently, political scientists have started using causal graphs Pearl (2009)—specifically, directed acyclic graphs (DAGs)—to communicate crucial assumptions more intuitively (Acharya, Blackwell, and Sen, 2016; Bellemare, Masaki, and Pepinsky, 2017; Blackwell and Glynn, 2018; Claassen, 2020; Imai and Kim, 2019; Keele, 2015b; Keele, Stevenson, and Elwert, 2020; Knox, Lowe, and Mummolo, 2020; Montgomery, Nyhan, and Torres, 2018; Samii, Paler, and Daly, 2016). However, the connection of causal graphs to the potential outcomes approach has remained unclear, at least to the uninitiated reader. To my knowledge, Imai and Kim (2019, fn. 5) is the first and only paper in political science to use both potential outcomes as well as a formal graphical argument, but does so only in passing.[2] This is at odds with developments in other disciplines such as epidemiology (Greenland, Pearl, and Robins, 1999), machine learning (Peters, Janzing, and Schölkopf, 2017), statistics (Maathuis and Colombo, 2015; Rothenhäusler, Ernest, Bühlmann, et al., 2018), and economic theory (Spiegler, 2016; Spiegler and Eliaz, Forthcoming), who have adopted DAGs more fully.

1 The earliest usage of counterfactuals using formal notation in political science seems to be Simon (1954).

2 Blair et al. (2019) rely on the notion of a causal model as in Pearl (2009), but do not associate it with a causal graph.

On the other hand, various researchers have stated that they generally find causal graphs to be not helpful Imbens and Rubin, 2015, p. 22.[3] And while Keele (2015b) explicitly uses DAGs for illustration purposes, he states that

> "in cases where identification conditions are well understood, a DAG may add little to the analysis. That is, in a well-conducted randomized experiment or a good natural experiment, the design creates such a simple DAG that they are of little use. However, under selection on observables, DAGs can be a useful way to clarify the necessary conditioning set".

Taken together, it seems that many researchers in political science have recently found causal graphs to be an intuitive and helpful visualization tool. However, they are reluctant to work with them as formal models, and they are pessimistic about their utility in analyzing simple and well-established research designs.

This article challenges this view and advocates that empirical political science and political methodology embrace insights from the literature on causal graphs (Pearl, 2009) more fully and explicitly. To this end, the article discusses formal tools that support main steps in causal analysis: Namely to derive, first, observable (testable) implications and, second, counterfactual (identification) assumptions, given a causal graph.

The article showcases the utility of DAGs by extending and clarifying a variety of methodologies that have been suggested recently. Specifically, it shows that, first, a statistical test for causal independence of mediators (Imai and Yamamoto, 2013) can be justified formally. Imai and Yamamoto (2013) state that this assumption, crucial for mediation analysis, cannot be tested, but nonetheless suggest to implement this test. A simple causal graph with four variables can be used to justify it, and also clarifies what a rejection of the Null tells us.

Second, the article shows that assumptions on unobserved confounders, as well as the causal independence of treatments in complex observational scenarios (Samii, Paler, and Daly, 2016) can be relaxed without costs. Here again, the underlying graph is of limited complexity. Also, a discussion of the

3 See Imbens 2019 for a more balanced assesssment.

graph in the context of the original application illuminates its substantive interpretation.

Third, the article proves that central assumptions on the independence of unobserved confounders and observed control variables in panel settings (Imai and Kim, 2019) can be similarly relaxed, which makes the use case of the associated methods considerably more realistic. Here, the graph and the involved arguments are considerably more complex, although the intuition can be deduced again from graphical identification criteria. This underscores the fact that identification in panel settings can be very challenging, and that careful application of formal graphical tools is needed to facilitate it.

Fourth, and finally, the article shows that the assumptions necessary for "compliance modeling" in instrumental-variables (IV) regression (Angrist and Fernandez-Val, 2013; Aronow and Carnegie, 2013; Esterling, Neblo, and Lazer, 2011) are, in the context of canonical IV graphs, equivalent to assuming no treatment-outcome confounding. This makes the use of IVs and compliance modeling as a related method superfluous. The focal graph again is simple (four observed variables). The challenge highlighted here, instead, is that assumptions on potential outcomes are hard to understand without aid by a graphical representation, and may imply parametric restrictions that are not obvious to the analyst.

While most of the tools discussed in this article are commonly used in machine learning (Peters, Janzing, and Schölkopf, 2017) and epidemiology (Greenland, Pearl, and Robins, 1999), this article also relies on a specific derivation of counterfactual assumptions more often used in econometrics (Imbens and Newey, 2009). In this regard, the article bridges some disciplinary divides. Furthermore, a concise discussion of the relationship between graphs and potential outcomes adds to otherwise fairly encompassing social science textbooks such as Morgan and Winship (2015).

Taken together, the analyses show that coupling DAGs and potential outcomes using formal rules allows us better understand and develop methodologies for causal inference. Hopefully, the article convinces the reader to take DAGs seriously, both as a set of substantive assumptions and as a formal tool to deduce the often non-obvious implications of such assumptions elegantly. Political science is at a unique position in this

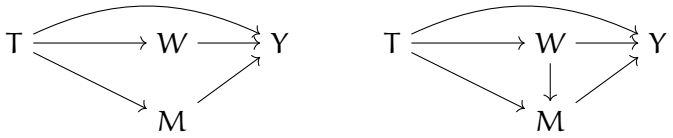

Figure 4: Causal graphs with two mediators, adapted from Imai and Yamamoto (2013). Left: Mediators do not influence each other. Right: W affects M and is a post-treatment confounder. Error terms not shown.

regard, because it has pushed the development of the potential outcomes framework, and has shown itself open to using causal graphs. In fact, Pearl made the prediction that "political science is destined to become a bastion of modern causal analysis" (Pearl, 2016).

The article does not discuss the basics of causal graphs separately but dives right into analyzing published examples. The next section illustrates the building blocks of DAGs using mediation analysis and introduces "d-separation" as a tool to derive testable implications. Section 2.2 briefly explain the "adjustment criterion", and then applies it to cross-sectional observational studies as well as panel analysis. Section 2.3 explains how graphs and potential outcomes are formally related and uses this to illuminate the assumptions behind compliance modeling. The final section concludes.

2.1 GRAPH BASICS, WITH AN APPLICATION TO MEDIATION ANALYSIS

We start by analyzing two simple graphs (Figure 4), adapted from Imai and Yamamoto (2013). In these two graphs, we have variables T, M, W, and Y. Each variable describes some distinct feature of a unit of observation. T is often taken to be the "treatment," i.e., the independent variable of interest. Y is usually taken to be the outcome of interest. Both graphs are "directed acyclic graphs" (DAGs). "Directed" indicates that every connection points one way or the other. "Acyclic" means that there are no cycles, i.e., no variable influences itself. For example, this means that if we added a connection $Y \rightarrow M$ to either of these graphs, they would cease to be acyclic, and most of what we discuss would not apply anymore.

We here interpret DAGs as "causal" graphs. Intuitively, a connection between two variables means that one affects the other causally, in a sense to be made more precise later. In general, the crucial assumptions are not the connections that are shown—these merely imply that there *could* be an effect—but the connections that are absent. In the context of Figure 4, this means that the second graph makes fewer assumptions than the first because it allows for an effect of W on M.

These graphs can be described as mediation models: T affects the outcome Y through the mediators M and W. Imai and Yamamoto (2013) discuss these graphs in the context of framing experiments, where T is a randomized frame shown to participants, and M, W, and Y are attitudes and beliefs measured post-treatment. Imai and Yamamoto (2013) show that in the first graph, where W and M do not affect each other, an encompassing mediation analysis is possible: One can assess the total effect of T on Y, its indirect effects through W and M, respectively, and its ("natural") direct effect. In the second graph, however, this is not possible. There, W acts as a "post-treatment confounder": A variable, observed or unobserved, that influences a mediator M of interest and the outcome Y (see also Acharya, Blackwell, and Sen, 2016).

Here, we focus on a more basic question: can we tell from the data whether we are in graph 1 or graph 2 (or in some other graph)? Put differently, given a causal graph, does it have an observable (testable) implication? For identifying mediation effects, this makes all the difference. This is discussed by Imai and Yamamoto (2013, p. 149), who note that "there exists no direct test of the assumed independence between causal mechanisms" (i.e., of W and M).[4] However, they still suggest to check for an association between the mediators conditional on treatment and pre-treatment controls. But without a formal justification, it is not clear how useful this test actually is.

In general, we can find such testable implications of a graph by enumerating all "paths" between two or more variables and checking whether they are "blocked", conditional on other variables M. Formally, a path is blocked if it involves a chain

4 This sentiment is shared by many other contributions to mediation analysis who either do not discuss testability of assumptions or state explicitly that they cannot be tested (e.g., Albert and Wang, 2015, p. 340; Caro, 2015, p. 584; Imai, Keele, and Yamamoto, 2010, p. 52; Keele, 2015a, p. 505; Naimi, Kaufman, and MacLehose, 2014, p. 1658).

$X \rightarrow M \rightarrow Y$ or a fork $X \leftarrow M \rightarrow Y$ and we condition on (control for) M, or a collider $X \rightarrow M \leftarrow Y$ and we do not condition on M.[5] To give some intuition, in the "chain" case, conditioning on M blocks the information flow from X to Y, in both directions. In the "fork" case, M acts as a "confounder" and creates a "spurious" association, but conditioning on it makes X and Y independent. Finally, in the "collider" case, learning the outcome M of two independent inputs X and Y enables us to predict the value of one input with the other.[6]

If some variables M block all paths between two sets of variables, we say that M "d-separates" these variables (in the graph) Pearl, 2009, p. 16.[7] Further, it follows that they are conditionally independent (in the data), for which the shorthand notation $X \perp\!\!\!\perp Y|M$ is used (Geiger, Verma, and Pearl, 1990). The software DAGitty (Textor, Hardt, and Knüppel, 2011), among others, automatically determines such independencies from a given graph. An easy-to-use `R` package is available, too (Textor et al., 2016).

In Figure 4, we see that in the first graph M and W are d-separated conditional on T. The first path $W \rightarrow Y \leftarrow M$ is blocked by Y acting as a collider. The second path $W \leftarrow T \rightarrow Y \leftarrow M$ is also blocked by Y. The third path $W \leftarrow T \rightarrow M$ is open, but can be blocked by conditioning on T, which is acting as a confounder. Taken together, this implies $M \perp\!\!\!\perp W|T$, which is exactly the association Imai and Yamamoto (2013) suggest to inspect (e.g.,by regressing M on W, T, and possible controls, and checking the coefficient on M). Furthermore, it is immediately clear that in the second graph, the testable implication does not hold, because W affects T directly so that they will generally correlate. However, there are other graphical structures imaginable that would rationalize a departure from the conditional independence implication. For example, there might be an unobserved confounder U of W and M (creating an open path $W \leftarrow U \rightarrow M$). However, an unobserved confounder of T and M or W, or of T and Y, would not lead to a rejection of the

5 Or any of M's descendants, i.e., any variable influenced by M.

6 See Acharya, Blackwell, and Sen (2016), Montgomery, Nyhan, and Torres (2018), and Knox, Lowe, and Mummolo (2020) for discussions of collider bias in empirical applications in political science, and Griffith et al. (2020) for possible collider phenomena in data on COVID-19.

7 d-separation stands for directional separation, as the criterion applies to directed graphs.

test (as it would not lead to open paths between W and M), even though it would invalidate identification of mediation effects. In this sense, the test is not a panacea for experimental design and theoretical assumptions. Appendix 2.A discusses this in more detail and also contains a more general proof of the validity of this testing strategy.

In general, it should be emphasized that given a graph, d-separation *logically* implies conditional independence in the absence of small-sample fluctuations. When one analyzes whether a certain quantity of interest—be it causal or non-causal—is "identified", one proceeds under this very same assumption of negligible randomness (Keele, 2015a). Finding a violation to a conditional independence relationship that is implied by a graph means that this graph cannot be the data-generating process.[8] However, it is not necessarily clear in which respect the graph is wrong, and furthermore, finding no violation does not mean the graph is appropriate. In this regard, it is important to make plausible assumptions (i.e., delete connections from the graphs only if one is sure that they do not exist), because they can only be tested in conjunction.

2.2 THE ADJUSTMENT CRITERION AND IDENTIFICATION

The prior section introduced a criterion—d-separation—that is very useful to understand whether certain variables are uncorrelated, given the causal structure visualized by the graph. We now take a step further and ask whether we can infer the magnitude of causal effects given a graph. In the potential outcomes literature, the canonical "ignorability" or "selection of observables" assumption $Y(t) \perp\!\!\!\perp T|X$ allows this. In words, this assumption states that the potential outcome of Y for some fixed value t of T is independent of the treatment T, conditional on other control variables X. If this and some further assumptions outside the focus of this article hold, then a matching or flexible regression estimator using X as control variables will consistently estimate some kind of average treatment effect of T on Y.

Imai and Kim (2019) is the first political science paper to refer to the "adjustment criterion" (Shpitser, VanderWeele, and

8 Save for, as usual, the possibility of a Type I-error.

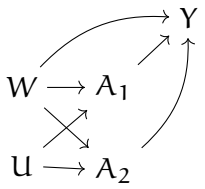

Figure 5: Slightly modified version of Figure 1 of Samii, Paler, and Daly (2016). Treatment of interest is (inter alia) A_1, outcome of interest Y. W are observed controls, U is unobserved. Allowing for additional effects of W on U (or vice versa) and of A_2 on A_1 would not hinder identification.

Robins, 2010) to justify such an ignorability assumption.[9] To identify the effect of T on Y using control variables, this criterion asks us to find variables X such that

1. all "non-causal" paths from T to Y are blocked by X and
2. no variable in X lies on or is influenced by a variable that lies on a causal path from T to Y.

A causal path is a path that goes away from T and enters Y. Condition 1 makes sure that we block all confounding paths between T and Y that create spurious associations. Condition 2 makes sure that we avoid post-treatment bias (Montgomery, Nyhan, and Torres, 2018). If we were to control for consequences of the treatment, then we might control away some part of the causal effect of interest, and also potentially introduce endogenous selection bias through a collider structure.

2.2.1 *The Adjustment Criterion in Observational Studies*

To see how the adjustment criterion can facilitate identification analysis, consider Figure 5, which is a slightly modified version of the causal graph in Samii, Paler, and Daly (2016). The paper analyzes observational data on demobilized Colombian guerrilla fighters and is interested in estimating causal effects to inform policy interventions to prevent recividism of these fighters. Here, W stands for observed control variables (gender, age, risk aversion, variables on fighters' military history, and more), $A = (A_1, A_2)$ are treatment variables (employment status and emotional well-being, among others), while U stands

9 The adjustment criterion is a generalization of the backdoor criterion developed by Pearl (1993).

for unobserved confounders that impact all variables in A, but, by assumption, not Y (recidivism). The interest is in the separate effects of all variables in A on Y. I will simplify matters a bit by concentrating only on the effect of A_1.

Samii (2016) do not discuss the substantive meaning of the variables in U (their focus is on developing and testing a new machine learning approach to estimating causal effects). For a more substantive-oriented research approach, this would seem highly relevant. For treatment variables employment status and emotional well-being, U would include variables that describe socialization experiences before participation in the war, some of which may be very hard to measure—e.g., to what extent an individual has had traumatic experiences. Furthermore, U may include variables describing the local economy that former fighters are residing in. The graph then assumes that none of these variables are related to W—neither through direct causal effects nor through shared common causes. As W is indeed a rich set of covariates, this seems exceedingly unlikely. For example, we would expect that early socialization affects individual education levels and military history, both of which are in W.

Furthermore, the graph assumes that none of the variables in U affect Y directly. This may be more or less plausible; here again, naming candidates for elements in U would be central.

Samii, Paler, and Daly (2016, p. 436) emphasize yet another assumption. They state that "an important assumption that this graph encodes is that, aside from the dependencies due to U and W, there are no direct causal relationships between the elements of A" (i.e., between A_1 and A_2). If such issues can be ruled out, the paper advocates to control for A_2 and W, and specifically suggests a flexible machine learning approach to do so.

Before I address this question directly, one should start analyzing this graph by asking whether it has a testable implication. The answer is "almost": W and U are clearly d-separated (unconditionally), but we do not measure U, so this is not testable. Furthermore, (A_1, A_2, W) d-separate U from Y, but this is similarly not testable. Finally, A_1 and A_2 will generally correlate because of the common confounder U, unlike in graph 1 of Figure 1. In sum, there is no testable implication, due to our inability to measure U.

Now, do A_2 and W satisfy the adjustment criterion for the effect of A_1 on Y, as implied by Samii, Paler, and Daly (2016)? The answer is yes: W blocks two non-causal path ($A_1 \leftarrow W \rightarrow Y$ and $A_1 \leftarrow W \rightarrow A_2 \rightarrow Y$), and A_2 blocks the third non-causal path $A_1 \leftarrow U \rightarrow A_2 \rightarrow Y$. Finally, neither A_2 nor W lie on a causal path from A_1 to Y.

However, we can greatly relax the assumptions in the graph without comprising identification. First, we can allow for either the $U \rightarrow W$ or the $W \rightarrow U$ effect (but not both, as this would create a cycle), plus common unobserved confounders (not shown in the graph). This introduces additional non-causal paths, but they are still blocked by W or A_2. Second, we can allow for an effect of A_2 on A_1.

Consider allowing for an $U \rightarrow W$ effect. This produces three additional non-causal path between A_1 and Y: $A1 \leftarrow U \leftarrow W \rightarrow Y$, $A_1 \leftarrow U \leftarrow W \rightarrow A_2 \rightarrow Y$ and $A_1 \leftarrow W \rightarrow U \rightarrow A_2 \rightarrow Y$. In all three paths, the observed W variables act as a confounder, and so conditioning on W blocks these paths. In sum, (W, A_2) still satisfy the adjustment criterion for the effect of A_1 on Y.[10]

This is fortunate because as suggested above, ruling out causal relationships between W and U or unobserved common causes seems implausible from a substantive point of view. However, allowing for $U \rightarrow Y$ indeed creates non-causal paths we cannot block, prohibiting identification through adjustment.

What happens if we allow for effects between A_1 and A_2, which Samii, Paler, and Daly (2016) mention to be problematic? There are two options. First, we may allow for an effect of A_1 on A_2. We see that now (A_2, W) would not satisfy the adjustment criterion anymore, as A_2 would lie on a causal path from A_1 to Y. In fact, we would now need to measure U directly, in order to block the non-causal path $A_1 \leftarrow U \rightarrow A_2 \rightarrow Y$ (which enters the outcome through a mediator). Therefore, identification fails.

However, allowing for the reverse impact of A_2 on A_1 does not impede identification. Starting from Figure 2 as it is, this would create new non-causal paths $A1 \leftarrow\!< -A2 \rightarrow Y$ and $A1 \leftarrow A2 \leftarrow W \rightarrow Y$. Both are blocked conditional on (A_2, W).

10 Additionally allowing for unobserved common causes of U and W, either via an explicit additional unobserved variable V or using the short-hand notation $W \leftrightarrow U$, is also not a problem. This adds non-causal paths $A_1 \leftarrow U \leftrightarrow W \rightarrow A_2 \rightarrow Y$, $A_1 \leftarrow U \leftrightarrow W \rightarrow Y$, and $A_1 \leftarrow W \leftrightarrow U \rightarrow A2 \rightarrow Y$. On all paths, W acts as a mediator between U and other variables, and conditioning on W blocks these paths.

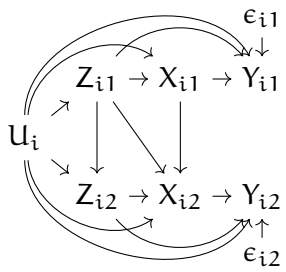

Figure 6: Causal graph for panel analysis with $T = 2$. Slightly modified version of Figure 3 of Imai and Kim (2019). Treatment of interest is now X_{it}. U_i are unobserved unit fixed-effects. Error terms ϵ_{it} are shown explicitly, and allowing for their impact on Z_{it} does not hinder identification.

In sum, this analysis shows that graphical identification criteria deliver a swift and constructive analysis of the impact of varying causal assumptions. It also highlights that analysts need to be careful when drawing or interpreting a graph. Researchers should try to name unobserved variables for treatment and outcome. Any absent arrow needs to be justified carefully; the same goes for assuming that a pair of variables does not share unobserved confounders.

2.2.2 *The Adjustment Criterion and Panel Analysis*

I now discuss how graphical identification criteria can improve the analysis of panel data with unobserved unit fixed-effects U_i, as in Figure 9. Here, X_{it} is the treatment for individual i at time t, Y_{it} is the outcome, and Z_{it} are time-varying control variables. I have also added time-varying error terms ϵ_{it} explicitly to the graph.

Imai and Kim (2019) show that in such a situation, a matching estimator can identify the average causal effect of X_{it} on Y_{it} for those units where X_{it} changes over time, which is an important generalization of identification in the classic linear fixed-effects model.

Interestingly, while the focus in Imai and Kim (2019) is on X_{it}, it might as well be on Z_{it}. In Figure 9, one can show that the same estimator using Z_{it} as the only independent variable—ignoring X_{it} altogether—identifies a corresponding average causal effect of Z_{it} (see Appendix 2.D). This indicates

that the assumptions in Figure 9 are even stronger than suggested by Imai and Kim (2019), because they allow for identification of causal effects of control variables. For example, Claassen (2020, Figure 2 c) uses a similar graph, where X_{it} is public support for democracy, Y_{it} is a measure of a nation's democratic status, and Z_{it} contains various other variables, such as GDP and resource dependence. Accordingly, the analysis seems to need to assume that there are no time-varying unobserved confounders of all of these control variables—not just X_{it}—and the outcome. But this is, of course, a very strong and generally implausible assumption (Keele, Stevenson, and Elwert, 2020), and would certainly need careful justification.

However, it is easy to show that if we allow for an impact of the unobserved variables ϵ_{it} on Z_{it}, the counterfactual assumptions needed for estimating the effect of X_{it}, but not of Z_{it}, are still fulfilled. Intuitively, we need to make sure that the Z_{it} and U_i variables fulfill the adjustment criterion for the effect of all X_{it} variables on all Y_it.[11] If we allow for effects like $\epsilon_{it} \rightarrow Z_{it}$, we introduce additional non-causal paths of the form $X_{it} \leftarrow Z_{it} \leftarrow \epsilon_{it} \rightarrow Y_{it}$. But these can all be blocked by conditioning on Z_{it}. On the other hand, if we were to concentrate on the effect of Z_{it}, we could never be able to block non-causal paths such as $Z_{it} \leftarrow \epsilon_{it} \rightarrow Y_{it}$. This once more underscores the asymmetric role of treatment and control variables.

2.3 THE STRUCTURAL DEFINITION OF COUNTERFACTUALS AND COMPLIANCE MODELING

d-separation gives all testable implications of a graph by looking at all paths between variables. The adjustment criterion and other graphical criteria allow for identification of certain causal effects, such as average treatment effects, controlled direct effects, and mediation effects, by directly checking whether certain paths (e.g., non-causal paths) can be blocked (see Appendix 2.A for a discussion of a graphical criterion for mediation analysis). But we might be interested in more general identification assumptions that are expressed using counterfactuals, for which graphical criteria do not (yet) exist. Also, we have not

11 Technically, one needs to rely on the structural definition of counterfactuals introduced in the next section. Appendix 2.D explains this in more detail.

seen directly how graphs and potential outcomes are formally related.

In the "structural causal model" framework Pearl (2009), this is achieved by treating DAGs as depictions of nonparametric structural equation models (NPSEM), and by defining counterfactuals using these structural equations. Thereby, for any counterfactual assumption, one can draw a graph that implies this assumption. It in this sense that the counterfactual and the "DAG framework" are equivalent (Pearl, 2009, p. 244). Imai and Kim (2019) were the first to rely explicitly on the NPSEM interpretation of DAGs in political science.

To see how this understanding of potential outcomes can improve causal analysis, consider "compliance modeling" for instrumental variables analysis (Angrist and Fernandez-Val, 2013; Aronow and Carnegie, 2013; Esterling, Neblo, and Lazer, 2011). Figure 7 shows a DAG for a prototypical scenario where Z is an instrument, U are unobserved confounders of treatment and outcome that make the application of the adjustment criterion impossible, and X are observed confounders.

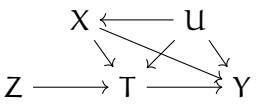

Figure 7: Z is a valid instrument for the effect of T on Y. Without further assumptions on the structural functions, the assumptions necessary for compliance modeling require to delete the $U \rightarrow T$ effect. U is unobserved. Error terms not shown.

The graph is equivalent to a set of structural equations as well as assumptions on the error terms. For example, it implies that

$$T = f(Z, X, U, \epsilon_T),$$

$$Y = g(T, X, U, \epsilon_Y),$$

$$\epsilon_T \perp\!\!\!\perp \epsilon_Y$$

Here, $f()$ and $g()$ are potentially very complex "data-generating processes" that for each unit measures all input variables—causes—and translates them deterministically into a value for T and Y. In principle, the errors ϵ play no special role. They are just variables that influence certain variables. Insofar as they are not observed, the observed variables are random,

i.e., they vary across units, even conditional on other observed variables. The independence assumption on the errors communicates the assumption that unobserved causes of the treatment do not affect the outcome and vice versa.

Under these assumptions, one can identify not the average treatment effect, but only the "local" average treatment effect for compliers (LATE). Formally, it is defined as $E[Y(T = 1) - Y(T = 0)|T(Z = 1) - T(Z = 0) = 1]$ in the binary instrument, binary treatment case. Compliance modeling (Angrist and Fernandez-Val, 2013; Aronow and Carnegie, 2013; Esterling, Neblo, and Lazer, 2011) tries to recover the often more interesting ATE parameter and suggests to look for observable control variables X such that "compliance status" $T(Z = 1) - T(Z = 0)$ is independent of the treatment effects of interest, $Y(T = 1) - Y(T = 0)$, that is, $T(Z = 1) - T(Z = 0) \perp\!\!\!\perp Y(T = 1) - Y(T = 0)|X$.

What does this assumption mean in the context of the graph in Figure 7? We cannot use the adjustment criterion, as it only allows us to derive the independence of observed treatment and potential outcomes. For example, it would hold that $Z \perp\!\!\!\perp Y(z)$ as well as $Z \perp\!\!\!\perp Y(z)|X$, as there are no open non-causal paths from the instrument Z to Y that need to be blocked, and additionally conditioning on X does not change this. $T \perp\!\!\!\perp Y(t)|X$ does not hold, because there are non-causal paths from T to Y running through U. But what about the assumption that $T(z) \perp\!\!\!\perp Y(t)|X$?

We can use the structural equations that lie behind the causal graph to define potential outcomes and to derive such complex counterfactual assumptions. The "structural definition of counterfactuals" defines them as values for variables when certain other variables are set externally to fixed constants in the respective structural equations Pearl, 2009, p. 204. This mirrors mathematically the thought experiment of a surgical, possibly counterfactual change of a variable. Accordingly, the potential outcomes would be

$$T(z) = f(z, X, U, \epsilon_T), \tag{30}$$

$$Y(t) = g(t, X, U, \epsilon_Y). \tag{31}$$

These expressions already utilize the fact that X and U are not influenced by Z or T, so that, for example, $X(t) = X$.[12] This definition of potential outcomes is actually often used in econometrics, albeit without reference to causal graphs (Chernozhukov et al., 2013; Imbens and Newey, 2009). De Mesquita and Tyson (2020) use a similar approach to defining potential outcomes. Finally, this definition makes clear that potential outcomes are a generalization of the concept of an error term. Appendix 2.B expands on this.

One can now taker a closer look at the assumption that compliance status is independent of treatment effects. Both are defined as differences in potential outcomes, but taking differences does not achieve much, since we do not know the form of f or g. So we might as well ask whether $Y(t) \perp\!\!\!\perp T(z)|X$ (which would imply the assumption needed). But looking at equations 30 and 31, it is apparent that this counterfactual implication is not true: U impacts both counterfactual variables, so they will be correlated, even if one adjusts for X. This is, of course, the original endogeneity problem in disguise.

The only way to make sure that compliance status is independent of treatment effects, in the context of this graph, is to rule out that U affects T. Then $T(z) = f(z, X, \epsilon_T)$ and $Y(t) = f(t, X, U, \epsilon_Y)$. Conditional on X, these variables are random only through ϵ_T and (U, ϵ_Y). Furthermore, conditional on X, ϵ_T and (U, ϵ_Y) are d-separated, so that $X(z) \perp\!\!\!\perp Y(d)|X$ holds (see Appendices 2.B and 2.E for more background on this derivation).

But if U does not affect T, we do not need an instrument (nor compliance modeling), as X also satisfies the adjustment criterion for the effect of T on Y. Accordingly, compliance modeling either implicitly assumes that there is no unobserved confounding, or it (again, implicitly) makes parametric assumptions that limit the extent of treatment effect heterogeneity.[13]

Consider the application in White (2019). White is interested in the effect of being sentenced to jail for misdemeanors (T) on voter turnout (Y), and shows that there is a negative effect for Black, but not for White citizens. She does so by instrumenting being sentenced by a measure of the strictness of a

12 By convention, unobserved error terms are always exogenous, i.e., not influenced by other variables.

13 Aronow and Carnegie (2013, p. 499) discuss the assumption that treatment effects are constant.

courtroom (Z) to which defendants are randomly assigned. She mentions the nature of the potential confounding variables (personal characteristics and offense severity) explicitly and makes a strong case for why the LATE is of scientific and normative interest in this context (White, 2019, pp. 316, 319). However, the ATE is of broader societal interest, as it informs us to what extent political demobilization due to the judicial system might affect election results.[14] Among other approaches, she uses compliance modeling to generalize her estimates, and finds that the ATE might be about twice as large as the LATE (White, 2019, Table A34). However, she can only do so using a very restricted set of covariates: Age, gender, severity of charge (class A or B misdemeanor), and past turnout. If we based the analysis of this identification approach on a DAG only, and not on stronger parametric assumptions, we would need to assume that these variables alone drive the confounding of sentencing and turnout, which seems implausible.

Formally, this section has proposed to derive counterfactual assumptions for which graphical criteria are not (yet) available using the structural definition of counterfactuals in conjunction with d-separation. That is, one needs to identify on which observed and unobserved variables counterfactuals depend, and then determine via d-separation whether these variables are (conditionally) independent of other variables. Appendix 2.B expands on this.

2.4 CONCLUSION

This article has introduced d-separation, the adjustment criterion, and the structural definition of counterfactuals in order to better understand and expand existing causal inference methodologies in political science, and to put published examples of DAGs into a broader methodological and substantive context. One bar was to show that DAGs can aid our understanding even in seemingly straightforward research designs with less than a handful of different variables.[15] By analyzing simple mediation and instrumental variables models and a

14 Although the average treatment effect on the treated may be even more informative for this.

15 Or more precisely, with less than a handful of different *roles* of variables, as usually X contains many control variables.

slightly more complex graph for observational studies, this article suggests that DAG can help here, too. For the case of panel analysis, the use case of DAGs seemed stronger a priori. Here, the article showed that graphical identification criteria lend intuition to determining whether a certain quantity is identified. Furthermore, in this case and the instrumental variables model that was analyzed, using the structural definition of counterfactuals in conjunction with d-separation seems indispensable (see Appendices 2.B and 2.D).

When graphs were analyzed in the context of actual research questions, it turned out that it is very helpful to name examples for important unobserved variables, and that one should not needlessly assume that these are unrelated to control variables. The fact that such assumptions commonly appear in published research may indicate that it is not widely known that the crucial assumptions in DAGs are the absent arrows.[16] However, one can be optimistic that researchers quickly adapt to the habit of looking for absent arrows.

There are interesting graphical approaches this article has not covered, especially with respect to deriving counterfactual assumptions. Among these are Single-World Intervention Graphs (Richardson and Robins, 2013), which may be useful for some purposes (see the discussion by Cinelli and Pearl 2018 and Appendix C). The approach suggested here uses the structural definition of counterfactuals in conjunction with d-separation. It is inspired by the practice in econometrics and can be used for all kinds of purposes.

It is common in both structural econometrics (Heckman and Pinto, 2015) as well as the potential outcomes literature (Imai and Yamamoto, 2013) to deduce further testable or counterfactual implications from given error or potential outcome assumptions using the conditional independency axioms due to Dawid (1979). Appendix 2.E discusses this approach and suggests that it is less transparent than using d-separation.

Empirical research of any kind is hard, and political science is nowadays at the forefront of developing new quantitative (Imai et al., 2011) and qualitative (Schneider, 2018) methodologies to solve research problems. For most of such methods, it is neces-

16 Relatedly, Keele, Stevenson, and Elwert (2020) document a widespread habit to give ill-justified (and needless) causal interpretations to the coefficients of control variables.

sary to communicate and understand complex assumptions. It is reassuring to see that political scientists have recently intuitively embraced causal graphs to aid this endeavor. Adopting them more fully should only lead to greater clarity, more exciting methodological developments, and interesting empirical insights.

APPENDIX

2.A IDENTIFICATION AND TESTABLE IMPLICATIONS OF MEDIATION MODELS

This section gives a formal proof that a graphical version of the Sequential Ignorability assumption in Imai, Keele, and Yamamoto (2010) (due to Pearl (2014)) implies a test if further measured mediators W are thought not to be influenced by the mediator of interest M. It also describes graphs that could give rise to a rejection of the test.

Definition. *Graphical Interpretation of Sequential Ignorability (Pearl, 2014)*
Assuming a given causal DAG G*, natural direct and indirect effect with respect to treatment* T*, mediator* M*, and outcome* Y *are identified if there exist measured covariates* X *such that*

1. X *and* T *block all* T*-avoiding backdoor paths from* M *to* Y *and*
2. X *blocks all backdoor paths from* T *to* M *and from* T *to* Y*, and no member of* X *is a descendant of* T*.*

I say that W is a *mediator* of the effect of T on Y whenever it is a descendant of T and an ancestor of Y. I then use the fact that d-separation rests on distinguishing only three basic types of paths, "chains", "forks", and "colliders". I also use the fact that when there is a collider path between two variables, the collider is a descendant of at least one of these variables. Then one can establish the following:

PROPOSITION When sequential ignorability holds on a causal DAG G for treatment T, mediator M, outcome Y, and control set X, any other vector of mediators W that is not influenced

by M will be d-separated from M, conditional on T and X. This implies

$$W \perp\!\!\!\perp M | D, X. \tag{32}$$

Proof. Assume the contrapositive; that is, in the graph, W and M are d-connected conditional on T and X, although sequential ignorability holds and M does not influence W. Then there must be at least one open path between W and M not blocked by X or T. There are three possibilities:

1) W influences M directly or they are connected through an unblocked "chain"

2) there is an unblocked backdoor path or "fork" between M and W not blocked by X or T

3) conditioning on X opens a path between M and W because it contains a collider or its descendant that is in X.

In the first case, there would be a T-avoiding backdoor path from M to Y over W not blocked by X, which would violate sequential ignorability. Same for the second case. In the third case, an element in X would need to be a descendant of M or W. In this case, by the definition of mediators, it would also be a descendant of T, which would violate sequential ignorability. □

This proposition makes precise and formally justifies the test advocated by Imai and Yamamoto (2013) and Loeys et al. (2013). The formulation of the proposition makes it clear that additional mediators influenced by the mediator of interest, M, are *not* a problem for sequential ignorability (see also Figure 6 in Imai et al. (2011)). Therefore, when analysts think such a variable exists, they can safely ignore it. But the proof suggests distinct (but mutually compatible) failures of the sequential ignorability assumptions in a graph that could lead to a dependence of M and W conditional on T. I will discuss each of these in turn. Note that in the above statement, W may be a vector containing multiple mediators, so any reasoning for one potential post-treatment confounders carries over to the case of multiple of such confounders. Specifically, the applicability and interpretation of the test hold regardless of the relationship between each of the potential post-treatment confounders, as long as cycles are ruled out.

Figure 8 gives three alternative DAGs that would lead to a rejection of $M \perp\!\!\!\perp W | T, X$, and that mirror the three cases men-

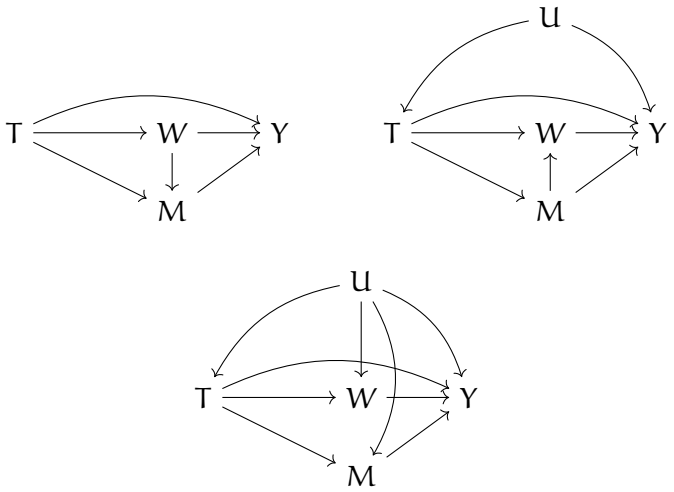

Figure 8: DAGs that would lead to a rejection of the testable implication of Proposition 1 and where Sequential Ignorability is violated. U is an unobserved confounder. Error terms not shown.

tioned in the proof. These should intuitively convey how informative the test of the implied conditional independence is. First, the test is not able to determine the direction of influence between the moderators. Both graphs on the top lead to a rejection of the independence of W and M, although the arrow between M and W is reversed in the latter. Secondly, the test is not able to detect confounding between T and Y (top right graph); this is also reflected in the fact that the proof of the proposition only considers T-avoiding backdoor paths. Lastly, the test can also not differentiate whether the association between M and W is causal or due to confounding, as in the bottom graph.

2.B THE STRUCTURAL DEFINITION OF COUNTERFACTUALS

Following Pearl (2009, p. 203), a causal model M has exogenous background variables (error terms), endogenous variables determined by other variables in the model, and structural functions. Each such model can be visualized by a directed graph (not necessarily acyclic): Nodes are variables, and there are links to a variable from each of the independent variables in its structural function.

A submodel M_x is a causal model where all functions for some variables X are replaced by constants x. This model gives the effect of the action $do(X = x)$. Furthermore, the potential

outcome of variables Y under this action is the solution for Y for the equations in M_x.

As an example, consider the graph $X \to Y$. The error terms are not shown. The associated causal model could be

$$X = f(\epsilon_X),$$

$$Y = g(X, \epsilon_Y),$$

$$\epsilon_X \perp\!\!\!\perp \epsilon_Y.$$

But we could also make a parametric assumption, and state that

$$Y = \alpha + \beta X + \epsilon_Y.$$

This is a textbook linear causal model. Here too, the structural definition of counterfactuals can be used. This leads to potential outcomes

$$Y(x) = \alpha + \beta x + \epsilon_Y.$$

Accordingly, both the individual and the average treatment effect would be β, a constant. Furthermore, the "ignorability" assumption $Y(x) \perp\!\!\!\perp X$ is equivalent to assuming $\epsilon_Y \perp\!\!\!\perp X$. This is because $Y(x)$ is random only as a function of ϵ_Y (see Appendix **??**). The graph tells us that this error term is d-separated from X, so that $\epsilon_Y \perp\!\!\!\perp X$. Accordingly, $Y(x)$, which is merely a linear function of the error term, is also independent of X.

More generally, potential outcomes can be functions of various observed and unobserved variables. Consider the simple mediation model $X \to M \to Y$. Here, the potential outcome $Y(x)$ is defined as

$$Y(x) = g(x, M(x), \epsilon_Y).$$

Here, x is a constant, but $M(x)$ is a random variable, itself a potential outcome, and defined as

$$M(x) = f(x, \epsilon_M).$$

Accordingly, ignorability would hold if X were independent of both $M(x)$ and ϵ_Y, because $Y(x)$ is merely a function of these random variables. X is independent of ϵ_M, and therefore of $M(x)$, and of ϵ_Y. So ignorability holds.

What happens when we condition on M, that is, does $X \perp\!\!\!\perp Y(x)|M$ also hold? For that to be the case, X would need to be d-separated from ϵ_M and ϵ_Y, as before. But if we add the error term ϵ_M to the graph explicitly, we notice the collider path $X \rightarrow M \leftarrow \epsilon_M$. Conditional on M, this path is open, X and ϵ_M are d-connected, and will generally correlate. This is another explanation for why the adjustment criterion prohibits to condition on variables on a causal path from treatment to outcome.

2.C SINGLE-WORLD INTERVENTION GRAPHS (SWIGS)

A single-world intervention graph (Richardson and Robins, 2013) makes the hypothetical intervention explicit in the graph, and splits the intervention variable into two variables. Thereby, some counterfactual independencies can be read off directly from the graph, using d-separation, without the necessity to commit to a representation using structural equations. For example, in the simple mediation graph from the previous section, an intervention $X = x$ would yield the SWIG

$$X|x \rightarrow M(x) \rightarrow Y(x).$$

The uppercase variable X is the same variable as in the original graph, whereas the lowercase x indicates the intervention, and observed variables M, Y become counterfactual variables $M(x), Y(x)$. The uppercase X variable inherits all incoming arrows (here: none), while the lowercase x inherits all outgoing arrows.

This way, it becomes clear immediately that under the simple mediation model, both $X \perp\!\!\!\perp Y(x)$ and $X \perp\!\!\!\perp Y(x)|M(x)$ hold, because X does not have any incoming or outgoing arrows in the SWIG, and so is d-separated from all other variables.

However, it is impossible to determine whether $X \perp\!\!\!\perp Y(x)|M$ holds using a SWIG in this example, which was easy using structural equations in the previous section. This is because

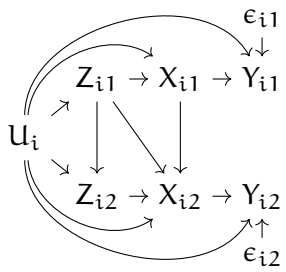

Figure 9: Causal graph for panel analysis with $T = 2$. Slightly modified version of Figure 3 of Imai and Kim (2019). U_i are unobserved unit fixed-effects.

once we consider the intervention $X = x$, the observed variable M becomes a potential outcome $M(x)$.

2.D IDENTIFICATION WITH PANEL DATA

Consider Figure 9, which is also in the main text, and is a slightly simplified (but substantially equivalent) version of Figure 3 in Imai and Kim (2019). Imai and Kim (2019) state that under the system of nonparametric structural equations associated with such a graph, an average causal effect of X_{it} on Y_{it} can be identified, by adjusting for measured Z_{it} directly and for unmeasured unit fixed-effects U_i indirectly. I now prove that under the same set of assumptions, a similar average causal effect of Z_{it} on Y_{it} can be identified, for which only adjustment for U_i is necessary.

Define potential outcomes $Y_{it}(Z_{it} = z)$ and define $C_i = I(0 < \sum_{t=1}^{T} Z_{it} < T)$. C_i is 1 for unit i if that unit experiences both treatment and control condition over the T time periods, i.e., it is a "switcher". The treatment effect of interest is $E[Y_{it}(Z_{it} = 1) - Y_{it}(Z_{it} = 0)|C_i = 1]$, the average effect of a binary Z_{it} for those units that switched treatment.

In line with Imai and Kim (2019), a sufficient assumption for identification would be

$$Y_{i1}(Z_{i1} = z), Y_{i2}(Z_{i2} = z)...Y_{it}(Z_{it} = z) \perp\!\!\!\perp Z_{i1}|U_i,$$

$$Y_{i1}(Z_{i1} = z), Y_{i2}(Z_{i2} = z)...Y_{it}(Z_{it} = z) \perp\!\!\!\perp Z_{i2}|U_i, Z_{i1}, \qquad (33)$$

...

$$Y_{i1}(Z_{i1} = z), Y_{i2}(Z_{i2} = z)...Y_{it}(Z_{it} = z) \perp\!\!\!\perp Z_{it} | U_i, Z_{i1}, Z_{i2}, ...Z_{it-1}.$$

These assumptions cannot be evaluated using the adjustment criterion. The adjustment criterion does not allow for checking, for example, whether $Y_{i2}(Z_{i2} = z) \perp\!\!\!\perp Z_{i1} | U_i$ holds, as the counterfactual is with respect to Z_{i2}, but the observed treatment variable is Z_{i1}. Instead, we need to resort to the structural definition of counterfactuals.

The relevant counterfactuals are all of the form

$$Y_{it}(Z_{it} = z) = g(z, X_{it}(Z_{it} = z), U_i, \epsilon^Y_{it}),$$

$$X_{it}(Z_{it} = z) = f(z, Z_{i1}, ..., Z_{it-1}, U_i, \epsilon^X_{it}).$$

Taken together, the counterfactual variable $Y_{it}(Z_{it} = z)$ is some function of random variables $Z_{i1}, ..., Z_{it-1}, U_i, \epsilon^X_{it}, \epsilon_{it}$. A collection of these counterfactuals as in assumption 33 is a function of error terms across all t.

Conditional on $(Z_{i1}, ..., Z_{it-1}, U_i)$, the randomness is only through the error terms $\epsilon^X_{i1}, ..., \epsilon^X_{it}, \epsilon_{i1}, ..., \epsilon_{it}$. Since conditional on $(Z_{i1}, ..., Z_{it-1}, U_i)$, Z_{it} is d-separated from these error terms, the counterfactual implications in 33 follows.

2.E DERIVING COUNTERFACTUAL AND TESTABLE IMPLICATIONS USING CONDITIONAL INDEPENDENCE AXIOMS

For random variables X, Y, Z, W, elementary properties of conditional independence are (Dawid, 1979; Pearl, 2009):

- Symmetry: $(X \perp\!\!\!\perp Y|Z) \implies (Y \perp\!\!\!\perp X|Z)$.
- Decomposition: $(X \perp\!\!\!\perp YW|Z) \implies (X \perp\!\!\!\perp Y|Z)$.
- Weak Union: $(X \perp\!\!\!\perp YW|Z) \implies (X \perp\!\!\!\perp Y|ZW)$.
- Contraction: $(X \perp\!\!\!\perp Y|Z)$ & $(X \perp\!\!\!\perp W|ZY) \implies (X \perp\!\!\!\perp YW|Z)$.
- Intersection: $(X \perp\!\!\!\perp W|ZY)$ & $(X \perp\!\!\!\perp Y|ZW) \implies (X \perp\!\!\!\perp YW|Z)$.

Furthermore, $X \perp\!\!\!\perp Y|Z$ is equivalent to $(X, Z) \perp\!\!\!\perp (Y, Z)|Z$. Additionally, if $X \perp\!\!\!\perp Y|Z$, and U is a function of X, then (i) $U \perp\!\!\!\perp Y|Z$ and (ii) $X \perp\!\!\!\perp Y|Z, U$ Dawid, 1979, Lemmas 4.1, 4.2. The latter property

may be called "transformation". The approach to derive counterfactuals from structural equations used in econometrics (e.g., Heckman and Pinto (2018)) crucially relies on this and the other conditional independence properties.

d-separation was proven to give all independence relationships (and only those) implied by a causal graph (Geiger, Verma, and Pearl, 1990), only relies on differentiating three different types of paths, and is automated in software packages (e.g., Textor, Hardt, and Knüppel, 2011). From this perspective alone, using conditional independence rules seems less attractive in terms of both rigor and transparency. This is further bolstered by the analysis below, which derives the testable implication of the mediation analysis as above using only structural equations and conditional independence rules. However, it should be pointed out that not all systems of relationships can be depicted as DAGs, and that there are independence relationships between variables that cannot be deduced using d-separation.

Using d-separation, one only needs to rely on transformation to derive counterfactual assumptions from a causal graph.

The left graph in Figure 4 in the main text implies the following structural equations and error independencies:

$$D = f_D(\epsilon_D), \tag{34}$$

$$W = f_W(D, \epsilon_W), \tag{35}$$

$$M = f_M(D, \epsilon_M), \tag{36}$$

$$\epsilon_D, \epsilon_W, \epsilon_M \text{ all jointly independent.} \tag{37}$$

We want to prove that $M \perp\!\!\!\perp W|D$.

Independence of error terms and transformation imply $T, W \perp\!\!\!\perp \epsilon_M$. Decomposition further implies $W \perp\!\!\!\perp T, \epsilon_M|T$. The structural equation for M then implies $W \perp\!\!\!\perp M|T$.

3

POST-INSTRUMENT BIAS

Co-authored with Adam Glynn and Miguel Rueda. See Part iv for author's contribution.

Identification of causal effects using instrumental variables is a popular approach in both experimental and observational research, and recent decades have seen an increasingly sophisticated understanding of what effects such instruments may identify. Based on the seminal work by Angrist, Imbens, and Rubin (1996), social scientists are nowadays aware of the role that assumptions such as the exclusion restriction or first-stage monotonicity play (Betz, Cook, and Hollenbach, 2018; Marshall, 2016; Sovey and Green, 2011). However, we contend that the choice of covariates in instrumental variable (IV) identification is not well-understood and leads to biases in applied research. Of special interest is the widespread adjustment for "post-instrument" variables to address a violation of the exclusion restriction, on which existing guidelines are either silent or contradictory. In this paper, we give straightforward advice for researchers on how to think about covariates in the context of IV analysis and for which of these one may need to adjust. To this end, we uncover significant new results and subtleties, especially with regards to (partial) tests of identifying assumptions. Furthermore, we develop a semi-parametric sensitivity analysis that aids applied researchers when there is a direct effect of an instrument that runs over measured variables.

Our contribution is motivated by both the widespread practice and voiced concerns of researchers that use instrumental variables. We have identified 116 papers published since 2010 in top political science journals[1] that use IV and explicitly discuss the exclusion restriction. Among those, one quarter (29 in total) use post-instrument covariates to justify the exclusion restriction. However, some researchers seem to be aware that adjustment for variables on other paths from instrument to outcome may not always lead to identification. For example, both

1 The American Political Science Review, the American Journal of Political Science, and the Journal of Politics.

Kern and Hainmueller (2009) and Carnegie and Marinov (2017) use instrumental variables and two-stage least-squares regression where they choose not (or not always) to control for such variables in order to avoid what they call "post-treatment bias". But there seems to be no justification for this in the literature, which uses this term for biases that are introduced in standard adjustment identification strategies, where instruments play no role (Angrist and Pischke, 2009; Montgomery, Nyhan, and Torres, 2018; Rosenbaum, 1984). On the other hand, Wucherpfennig, Hunziker, and Cederman (2016) claim that "the instrumental variable logic is immune to any correlation (and even causation) between the instruments and the covariates". This position actually finds support in a leading econometrics textbook (Wooldridge, 2010, pp. 94, 938). Other standard textbooks like Angrist and Pischke (2009) and reader's guides like Sovey and Green (2011) are silent on such issues. The need to formally discuss the role of covariates in instrumental variables analysis is also echoed by Lee and Lemieux (2010), who observe that "it is often unclear which covariates to include in the analysis", and that adjustment for more variables may not always be desirable. However, they also give no clear advice.

To fix ideas, consider an example from Angrist (1990), whose identification strategy has inspired several studies of political behavior (see Berinsky and Chatfield 2015 for an overview). The author is interested in estimating the effect of serving in the Vietnam war on earnings. The draft was largely determined by a randomized lottery, and Angrist notes that men who have a low draft lottery number were more likely to serve in the war. He uses functions of this number as instruments for military service.

There could be some concerns about the validity of the exclusion restriction. For example, those who received a low lottery number could have chosen to stay in school to obtain a deferment (Angrist, 1990, p. 330). This creates a link between the lottery and earnings via education. So if the information on post-lottery education was available, should we control for it?

In this paper, we answer this question and discuss various related problems. We rely on the framework of "structural causal models" that unifies both potential outcome and graphical approaches to causality (Pearl, 2009). This allows us to give advice to applied researchers that is both easy to formulate and under-

stand. We first make clear the asymmetric role of pre- and post-instrumental variables. Then, we illustrate how adjustment for variables influenced by the instrument may not always be successful, and that adjustment for variables influenced by the *treatment* will lead to biases in IV identification even when the IV is unconditionally valid. The mechanics behind these phenomena resemble the better-known "post-treatment" bias in adjustment strategies (Montgomery, Nyhan, and Torres, 2018), although additional, more subtle problems occur. However, we also show, perhaps to the surprise of some researchers, that adjustment for variables influenced by the *instrument* is sometimes *necessary* for successful identification. In some cases, we show that this identifies the well-known local or weighted average treatment effect. For other cases, we propose to identify a new, different treatment effect. In sum, "post-instrument bias" is quite different from "post-treatment bias".

The assumptions for valid post-instrument adjustment are highly restrictive, although we also prove that they are testable under some circumstances. In this context, we discuss the evidential value and implicit causal assumptions of other informal tests and robustness checks that are prevalent in the applied literature. We show that these tests are generally *misleading*. Therefore, we also add to the theory of robustness checks, which so far has concentrated on regression adjustment strategies (Chen and Pearl, 2015; Lu and White, 2014).

What if the strong assumption necessary for identification are not plausible or rejected by the data? We propose that researchers utilize measures of the variable on the pathway from instrument to outcome for a semi-parametric sensitivity analysis. Our approach generalizes previous approaches (Conley, Hansen, and Rossi, 2012; Van Kippersluis and Rietveld, 2018) that operate under a strong effect homogeneity assumption and cannot use sample information to bound biases. Moreover, our approach also works if there is a measurement error in the post-instrument variable. This will often be the case when potential violations of the exclusion restriction are uncovered only after initial data collection and intense scrutiny of an IV strategy. We illustrate our approach by reanalyzing the data of Spenkuch and Tillmann (2017) on the causal effect of Catholicism on the Nazi vote share at the end of the Weimar republic. The application highlights the need to relax stringent linearity assump-

tions and to account for potential heterogeneity in causal effects. However, the main inference of this paper appears to be robust to all but extreme heterogeneity.

A related paper is Deuchert and Huber (2017), who point out that investigating instruments that may affect more than one variable is also highly relevant because oftentimes the same instrument is used to study causal effects of different treatment variables, so that researchers might be tempted to blindly adjust for these other treatments. For example, Bazzi and Clemens (2013) discuss the "origin of a country's legal system" instrument that has been used for at least seven different treatments. Similar to our approach, Deuchert and Huber (2017) also use causal graphs. However, they use these for illustrative purposes only and prove their main results under a strong linearity assumption. In contrast, we discuss these issues in a completely nonparametric framework and integrate causal graphs with the popular potential outcomes approach. Importantly, we discuss additional identification assumptions, prove that these are sometimes testable, introduce a new causal estimand, and propose a new sensitivity analysis. We also correct a mistaken statement in Deuchert and Huber (2017) regarding a central identification problem. Betz, Cook, and Hollenbach (2018) also use causal graphs to illustrate failures of IV identification using "spatial" instruments. Finally, many of the problems that we discuss are similar to what Elwert and Winship (2014b) call "endogenous selection bias". They focused on classic treatment models where instruments are not available, so that our paper naturally complements their analysis.

3.1 UNDERSTANDING CONDITIONAL IV IDENTIFICATION USING CAUSAL GRAPHS

In this section, we first give an introduction to the formal tools we use to tackle the problem of covariate adjustment in IV identification. We then present a series of causal graphs that allow for identification of various treatment effects when the key "ignorability" assumption only holds conditionally. We use causal graphs because they offer a straightforward formalization of the language already used by most researchers to communicate assumptions about the causal ordering of variables, direct and indirect effects, confounding, etc. Additionally, they can

be integrated with the popular potential outcomes approach to causality, and allow for a derivation of assumptions on the distribution of these potential outcomes. A formulation of the same assumptions using counterfactuals only is possible, but much more tedious and intransparent, see for example Pearl (2009, pp. 128–132). For other recent uses of causal graphs in political science, see Imai et al. (2011), Imai and Yamamoto (2013), and Glynn and Kashin (2017).

Consider again our example from Angrist's seminal analysis Angrist (1990). Angrist is interested in the causal effect of serving as a soldier in the Vietnam war (D_i) on later earnings Y_i. The draft lottery leads to a binary instrument Z_i that indicates draft eligibility.

The "ceiling" for the draft varied by year due to fluctuating demands by the military. Therefore, the cohort X_i of a man influenced the probability that he would be drafted. At the same time, birth year is clearly causally prior to the draft and might have other effects on the outcome. This can easily be depicted in a causal graph such as figure 10.

The dashed arrows emanating from the U_i-variable indicate that it stands for unobserved variables that may influence treatment, outcome, and covariates X_i, but not the instrument. In the Vietnam draft example, U_i may contain variables describing the socio-economic status of one's parents. These will impact on the decision to enlist in the military, and on later socio-economic outcomes. They may also affect the timing of birth. The existence of such unobserved confounders is the central motivation for employing IV identification, because they make identification of the effect of D_i on Y_i via regression or matching impossible. With this first example in mind, we now discuss how to formally derive identification assumptions from causal graphs.

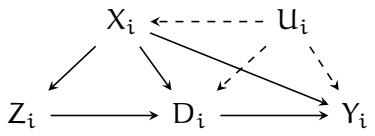

Figure 10: Benchmark graph. In this graph, Z_i is an instrument for the effect of D_i on Y_i conditional on X_i, but not unconditionally.

3.1.1 *Causal Graphs and d-Separation*

Causal graphs, specifically *directed acyclic graphs*, consist of *nodes*, which visualize variables, and *edges*, which are usually directed arrows from one node to another. A *path* is any consecutive sequence of edges. In line with Pearl (2009), we view causal graphs as depictions of a nonparametric system of structural equations that describes cause-effect relationships. That is, nodes stand for observable or unobservable features of the units of interests, and an edge or arrow from one such node to the other communicates the assumption that the one variable causally affects the other variable in the population of interest. To be precise, a causal model G consists of exogenous background variables U_i, usually assumed to be unobserved, observed endogenous[2] variables V_i, and structural (causal) functions f_v for each endogenous variable. These functions are deterministic in the sense that if we knew all relevant inputs of f_v for an endogenous variable, we could determine the value of this variable exactly. Since U_i is assumed to be unknown, the observable variables V_i become random variables. Whenever we want to indicate that observable variables are driven by an unobserved confounder, we will use dashed nodes for edges emanating from this confounder. This is equivalent to assuming that the "structural errors" U_i (i.e., all unobserved causes) of the confounded variables are dependent. Throughout, we discuss *acyclic* graphs, that is, graphs where no variable may have an effect on itself. Finally, we use upper-case letters do denote random variables, and lower-case letters to denote realized or fixed values of these variables.

To understand in which situations an instrument is (conditionally) valid, it is necessary to derive independence relationships from the causal graph the researcher assumes. Throughout, we do so by using an easy yet powerful tool called *d-separation* (Geiger, Verma, and Pearl, 1990). In a given graph, a path p is said to be d-separated (or *blocked*) by a set of nodes Z_i if and only if

1. p contains a chain $X_i \to M_i \to Y_i$ or a fork $X_i \leftarrow M_i \to Y_i$ such that the middle node M_i is in Z_i, or

2 Here, the word "endogenous" simply means "explained in the model".

2. p contains an inverted fork (or *collider*) $X_i \rightarrow M_i \leftarrow Y_i$ such that the middle node M_i is not in Z_i and such that no descendant of M_i is in Z_i.

A set of variables Z_i is then said to d-separate X_i from Y_i if and only if Z_i blocks every path from a node in X_i to a node in Y_i. Importantly, d-separation implies conditional independence, which we write as $X_i \perp\!\!\!\perp Y_i | Z_i$. This means that, once we know the value of Z_i, X_i does not predict Y_i and vice versa. In addition, we employ graphoid axioms (Dawid, 1979) to prove our results. We expand on these more technical aspects and give proofs in the appendix. In the main body of this article, we stick as closely as possible to intuitive explanations.

The fact that conditioning on a collider of two variables (or its descendant) makes these variables dependent is central to understanding the failure of certain IV strategies, but may be counterintuitive, so that an example is helpful. Consider two independent binary variables A and B and a random variable C that is the sum of A and B. Accordingly, C can take on the values $\{0, 1, 2\}$, and is a collider variable, with A and B pointing into it. A and B may be random coin flips, so clearly knowing the value of A does not help in predicting B. However, conditioning on the collider C means that we are told its value, for example, 1. The question then is whether A and B have become dependent, that is, whether knowing C and A now tells us anything about B. The answer is a clear yes: Knowing the result C is 1 and, for example, that A is 0, we know for sure that B has to be 1. Put differently, knowing the result of a process (C) and the value of one of its independent inputs (A) also lets us predict the value of the other input (B). The same mechanics apply if we happen to know the realization of a descendant of C. For example, let D be a variable that takes on the value 1 when C equals 1, and is 0 otherwise (so that it is a binary proxy for C). Knowing that D equals 1 and that A equals 0 also leads to the prediction that B equals 1.

To give a more elaborate example of d-separation, consider figure 10. Let us assume for the moment that we could measure U_i and we were interested in its dependency with Z_i . In this case, one would find four paths between the instrument Z_i and U_i: $Z_i \rightarrow D_i \leftarrow U_i$, $Z_i \leftarrow X_i \rightarrow D_i \leftarrow U_i$, $Z_i \rightarrow D_i \rightarrow Y_i \leftarrow U_i$, and $Z_i \leftarrow X_i \rightarrow Y_i \leftarrow U_i$. The first two paths contain the variable D_i as a collider and so are unconditionally blocked. The

last two paths contain Y_i as a collider and therefore are blocked as well. In summary, all paths between Z_i and U_i are blocked unconditionally, so that Z_i and U_i are d-separated and $Z_i \perp\!\!\!\perp U_i$ holds. Put informally, this conveys the notion that a valid instrument needs to be independent from unmeasured causes of Y_i. Accordingly, if one could measure U_i for each individual, a linear regression of it on Z_i should yield a coefficient of zero (asymptotically).

3.1.2 *From Graphs to Potential Outcomes*

Having discussed the basic properties of causal graphs, we now introduce potential outcomes and the causal effects of interests. As usual, the identification assumptions need to be stated as independence relationships of observed and counterfactual variables. Following Pearl (2009), we connect causal graphs and potential outcomes by defining the latter quite naturally as solutions to the structural model that researchers assume. The potential outcome of variables $Y_i \in V_i$ when variables $X_i \in V_i$ are set to x is denoted $Y_i(X = x)$ and is given by $Y_i(G_x)$. G_x stands for a manipulated version of the original causal model G in which all functions f_{X_i} are deleted and replaced by constants x Pearl, 2009, p. 204.

To give a simple example, consider the graph $D_i \rightarrow Y_i \leftarrow U_i$. In this graph, the potential outcome of Y_i in unit i when D_i is set to d is

$$Y_i(D = d) = f_y(d, U_i)$$

which, since d is fixed, is a random variable only because it is a function U_i, which stands for all unobserved causes of Y_i. It follows immediately that $D_i \perp\!\!\!\perp Y_i(D_i = d)$ ("ignorability") holds, because D_i and U_i are d-separated unconditionally (since Y_i is a collider that blocks the only path between D_i and U_i). In DAGs, ignorability of the treatment can also be evaluated by simple graphical criteria like the adjustment criterion (Shpitser, VanderWeele, and Robins, 2010). However, we resort to this structural definition of counterfactuals to make explicit the exact reasons for why IV identification may fail, and because such general graphical criteria for IV problems do not exist.

Our approach is fully compatible with previous results that used counterfactuals to communicate causal assumptions. Ap-

proaches that define potential outcomes as byproducts of a structural equation are also becoming standard in econometrics, see for example Imbens and Newey (2009), Chernozhukov et al. (2013), and especially White and Lu (2011a), who also employ causal graphs. It should also become clear that potential outcomes are indeed a generalization and refinement of the "structural error" that plays a central role in econometrics. Again, this error term in a structural or causal equation stands for all unobserved factors that influence the outcome when observed determinants are held fixed, and it should not be confused with the regression error. The latter stands for a unit's deviations in Y_i from its conditional mean. See Imbens (2014) for a discussion of this issue in an IV context.

Generally, we will discuss identification of variants of a local average treatment effect (LATE):

$$E[Y_i(D = 1) - Y_i(D = 0)|D_i(Z = 1) > D_i(Z = 0), X_i]$$

This is the average causal effect of a binary treatment D_i on outcome Y_i among those individuals 1) for which an instrument Z_i changes treatment status (compliers) and 2) which are characterized by covariate profile X_i. What if the treatment is continuous? First write the causal effect of instrument on treatment as $D_i(Z = 1) > D_i(Z = 0) = \alpha_i$. If the structural equation of interest has heterogeneous effects, but otherwise is linear, as in

$$Y_i = \mu_Y + \beta_i D_i + \epsilon_i,$$

then the parameter of interest is usually

$$\frac{E[\alpha_i \beta_i]}{E[\alpha_i]} = E\left[\frac{\alpha_i}{E[\alpha_i]}\beta_i\right]. \tag{38}$$

Conventionally, three assumptions are used to identify such treatment effects. These are often discussed for the case of binary instrument and treatment, although they easily generalize. The first assumption, monotonicity, assumes that

$$P(D_i(Z_i = 1|X_i)) \geqslant P(D_i(Z_i = 0|X_i)) = 1$$

That is, the instrument has a causal effect on the treatment that pushes every unit in the same directions, and there are no "defiers". If this holds, $\alpha_i \geqslant 0$, so that the expression in equation 38

is a weighted average of individual-treatment effects β_i, where the weights are all greater than or equal to zero.

Secondly, it is assumed that Z_i and D_i are dependent ("relevance"):

$$E[D_i|Z_i = 1, X_i] - E[D_i|Z_i = 0, X_i] \neq 0$$

which is directly testable. In this paper, we will focus on understanding the crucial conditional independence assumption (CIA)

$$Y_i(D_i = d), D_i(Z_i = z) \perp\!\!\!\perp Z_i|X_i$$

In words, this assumptions states that the potential outcome of outcome Y_i when treatment D_i is set to d and the potential outcome of D_i when instrument Z_i is set to z are jointly independent from Z_i, given covariates X_i.

If these assumptions—CIA, monotonicity, and relevance—hold, two-stage least squares with saturated models in both stages estimates a weighted average of these X_i-specific LATEs, and this approach is dominant in applied research (Angrist and Imbens, 1995; Angrist and Pischke, 2009, p. 177). Notably, the CIA subsumes both the exclusion restriction and the more opaque "ignorability" requirement. We use graphs to illustrate when this latter assumption hold, and will usually discuss the "causal first-stage" assumption $D_i(Z = z) \perp\!\!\!\perp Z_i|X_i$ separately from the $Y_i(D_i = d) \perp\!\!\!\perp Z_i|X_i$ requirement, since this is more intuitive. Formal derivations of the joint independence are relegated to the appendix.

3.1.3 *Identification with Pre-Instrument Covariates*

We start with a benchmark graph (figure 10). In this graph, the treatment and outcome are driven by unobserved confounders U_i, while there are also observed confounders X_i that may influence the instrument, treatment, and outcome. This will not be the case when Z_i is physically and unconditionally randomized, because this precludes the $X_i \rightarrow Z_i$ path. However, if there are such observed confounders, adjustment for them is necessary. Intuitively, a first-stage regression of D_i on Z_i only would not give the causal effect of Z_i on D_i because of the open "backdoor" paths $Z_i \leftarrow X_i \rightarrow D_i$ and $Z_i \leftarrow X_i \leftarrow U_i \rightarrow D_i$. Similarly, the instrument and the outcome would be connected through a

path other than the effect going through D_i. Conditioning on X_i solves both problems, because X_i blocks these spurious paths.

Put formally, both $D_i(Z_i = z)$ and $Y_i(D_i = d)$ are a function of random variables U_i and X_i:

$$D_i(Z_i = z) = f_d(z, X_i, U_i)$$

$$Y_i(D_i = d) = f_y(d, X_i, U_i)$$

Conditional on X_i, these counterfactuals are random only through U_i, and we see that conditional on X_i, Z_i and U_i are d-separated. Accordingly, the CIA holds.

The CIA does not hold when two key conditions are violated. First, it may be that the confounders U_i also influence the instrument Z_i; in this case, Z_i and U_i are dependent (d-connected), and conditioning on X_i does not break this dependence. This is the problem of "backdoor paths" which has found extensive treatment in the graphical literature. In fact, it suffices to have unobserved confounders that influence Z_i and D_i or Z_i and Y_i, but not necessarily all three variables, to invalidate the CIA. This fact has been overlooked even by very careful applied researchers that made their thinking about potential confounders explicit (e.g. Kocher, Pepinsky, and Kalyvas 2011, p. 212; Stanig 2015, p. 188).

Second, Z_i may have an effect on Y_i going not through D_i. In this case,

$$Y_i(D_i = d) = f_y(d, Z_i, X_i, U_i)$$

which clearly depends on Z_i, so that the CIA is violated.[3]. Note, howevever, that $D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i$ may still hold if the exclusion restriction is violated, so that the effect of Z_i on D_i remains identified.

In the following, we will assume that observed pre-instrument covariates X_i may exist, and that conditioning on them solves the "backdoor" problem. Specifically, this will even hold if U_i influences X_i (so that the effects of variables in X_i are not identified). This relaxes the common restriction for all X_i variables to be "exogenous" (e.g. Wooldridge 2010, p. 110), and

3 Accordingly, the exclusion restriction is part of the CIA in our formulation, a point also made by Angrist and Pischke (2009, p. 132) In their guide to IV, Sovey and Green (2011, p. 191) state that the exclusion restriction is needed in addition to this CIA, but this is actually not true. It becomes clear when one thinks of potential outcomes as a summary of variables that influence Y_i when D_i if physically fixed.

differentiates such control variables from the post-instrument variables we discuss next. For ease of visual presentation, we will not depict the X_i nodes in the causal graphs that we discuss in the remainder of this article.

3.1.4 *Identification when Covariates are Influenced by the Treatment*

We now discuss a variety of situations in which researchers measure covariates M_i that are influenced by the instrument, that influence the outcome, and that may also influence or be influenced by the treatment.[4] Our main result is that identification of a local average treatment effect is possible in some cases under strong assumptions. It turns out that identification relies on adjustment for the M_i covariates, even if they also influence the treatment. For the latter case, we introduce a new causal estimand and show how it is identified. Accordingly, "post-instrument" bias does not generally occur, but depends on the causal model. Additionally, ruling out causation between D_i and M_i allows for a test of the identification assumptions which is easy to implement. We discuss other, informal tests in the literature and show that these are generally misleading.

In the Vietnam draft example, a potential M_i variables is college education, because the latter may have been used to avoid the draft, and because it plausibly affects earnings. The textbook by Wooldridge (2010, p. 938) discusses this complication and claims that statistical adjustment for such a variable "effectively solves this problem". In the following, we show that this statement needs considerable qualification.

The most simple case is shown in graph (a) in figure 11, where the variable M_i is influenced by the instrument Z_i and in turn is a cause of Y_i. However, neither does D_i drive M_i, nor does M_i influence D_i, nor is U_i influencing M_i. Can we then simply control for the "post-instrument" variable M_i to make the instrumental variable approach work? Such an explicit approach is chose, inter alia, by Abramson (2017), Ahmed (2012), Boix (2011), Pierskalla and Hollenbach (2013), Spenkuch and Tillmann (2017), Trounstine (2016), and Woodberry (2012).

It turns that under the restrictive assumptions visualized in graph (a), this conditioning strategy indeed identifies an

4 Our results only hold for *acyclic* graphs. This means that researchers need to rule out mutual causality between variables a priori.

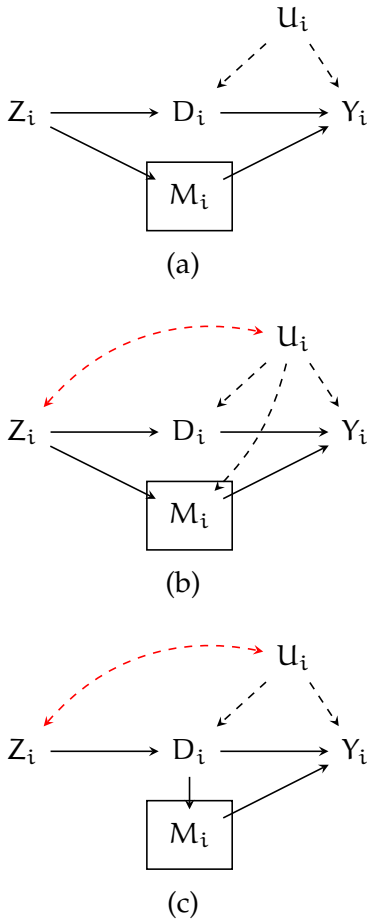

Figure 11: Three prototypical IV scenarios with post-instrument variables. Boxes indicate the conditioning on M_i, and red bi-directed arrows indicate dependencies created by such conditioning. In graph (a), conditioning on M_i is required and identifies the M_i-specific local effect of D_i on Y_i. In graph (b), conditioning on the collider M_i opens a non-causal path between U_i and Z_i. In graph (c), M_i is a descendant of collider D_i, and the same dependence by Z_i and U_i is created.

(X_i, M_i)-specific LATE or weighted ATE as in equation 38, since the CIA holds with conditioning set (X_i, M_i). To see why, consider the first-stage effect of Z_i on D_i. Although M_i is "post-instrument" - i.e., influenced by Z_i - conditioning on it does not invalidate the ignorability of Z_i with regards to D_i, i.e. $D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i, M_i$ holds. This follows from the graphical adjustment criterion (Shpitser, VanderWeele, and Robins, 2010). For an alternative explanation, note that under this graph

$$D_i(Z_i = z) = f_d(z, X_i, U_i)$$

Since Z_i is independent from U_i conditional on X_i and M_i (by d-separation), Z_i is also conditionally independent from this potential outcome. Intuitively, there is no "backdoor" path from Z_i to D_i not blocked by X_i, and conditioning on M_i does not block any genuinely causal paths, nor does it open up any new spurious paths, since it is not a collider. In a similar vein,

$$Y_i(D_i = d) = f_y(d, M_i, X_i, U_i)$$

is independent from Z_i conditional on M_i and X_i, because the direct path through M_i is blocked while no other paths are opened up.

There are two crucial assumptions for the validity of this approach that may be violated. First, it may be that M_i is also driven by the unobserved confounder U_i. This situation is depicted in graph (b) of figure 11. In our running example, it is quite easy to imagine that unobserved parental SES positively influences the choice to go to college directly. In this case, M_i becomes a collider, and conditioning on it (indicated by the box around it) opens up an unblockable path (indicated by the dashed by-directed arrow) between Z_i and U_i. Specifically, we would compare draftees ($Z = 1$) to non-draftees ($Z = 0$), given the same college decision $M_i = m$. If Z actually affects the college decision, then the fact that the latter is observed to be constant in such a group must be due to individual differences in U_i, which then affect Y_i irrespective of an actual treatment effect. E.g., draftees that did not attend college to avoid the draft probably had lower parental SES than non-draftees, and lower wages Y_i for that reason alone - even if neither treatment nor college actually affected earnings.

This open "non-causal" path then actually invalidates both the first-stage $D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i, M_i$ assumption due to post-treatment selection bias,[5] as well as the $Y_i(D_i = d) \perp\!\!\!\perp Z_i | X_i, M_i$ assumption.

Second, even if Z_i does not *directly* drive M_i, the latter may be influenced by the treatment D_i, as in graph (c) of figure 11. In this case, M_i is a mediator of the $D_i \rightarrow Y_i$ relationship, and is also influenced by Z_i indirectly through D_i. Wooldridge (2010, p. 95) suggests that on-the-job training might be such a variable in the Vietnam draft application. In this case, Z_i is a

5 For an in-depth analysis of this phenomenon in standard adjustment strategies in political science, see Montgomery, Nyhan, and Torres (2018).

valid instrument when one does *not* adjust for M_i. This is because the exclusion restriction obviously holds, and there are also no other backdoor paths that connect Z_i and Y_i. However, adjusting for M_i introduces a severe, but more subtle problem. D-separation does not only prohibit to condition on colliders to block paths, but also to condition on *descendants* of such variables. Since Z_i and U_i collide in D_i, conditioning on its "child" M_i has the same qualitative consequences as in graph (b), making it impossible to identify the ATE of Z_i on D_i or the LATE of D_i on Y_i. This subtle problem went unnoticed by Deuchert and Huber (2017, p. 416), who discuss a similar graph and state that conditioning on a mediator satisfies the CIA and identifies a "partial direct effect". As we hope we have made clear, this is not the case, because conditioning on a mediator renders Z_i correlated with U_i, which prohibits any identification.[6] We return to these graphs again when we discuss the possibility of testing which of the assumptions hold.

An interesting special case of graph (c) of figure 11 is when M_i stands for the inclusion of an observation in the dataset (or, reversely, for attrition). In both observational and experimental studies, participants often drop out based on the realization of their treatment (Elwert and Winship, 2014b, p. 42). Researchers are then forced to condition on M_i. In IV settings, even if M_i is not directly driven by U_i and does not influence Y_i, it is a descendant of the collider D_i, so that the instrumental variable becomes invalid. Similarly, in Angrist (1990), it is noted that reported earnings are censored at a maximum l, so that the whole sample is conditional on $Y \leqslant l$. This means one conditions on a descendant of the true unobserved earnings, so that the IV becomes invalid, a fact acknowledged by Angrist (1990, p. 334). Berinsky and Chatfield (2015) discuss this and related selection problems that may occur for the draft of the lottery instrument.

A final possible set of causal assumptions is depicted in graph 12. In this graph, M_i is not influenced by the confounder U_i, but affects D_i. Again, the no-confounding assumption is crucial. If it is violated, a collider phenomenon would occur as in the previous cases, making Z_i an invalid instrument. How-

6 Frölich and Huber (2017) propose to identify mediation effects in such a setting using an instrument influencing D_i and a separate instrument influencing M_i.

ever, if such confounding can be ruled out, one can actually identify a local ATE:

$$E[Y_i(D_i = 1) - Y_i(D_i = 0) | D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m), X_i]$$

This estimand has not been discussed before. It is the average causal effect of a binary treatment for the latent subpopulation of units which 1) change treatment status as a response to the instrument Z_i, *while fixing* M_i *at* m and 2) which are characterized by covariates X_i.

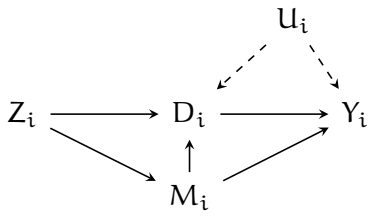

Figure 12: Graph where adjustment for M_i identifies a local average treatment effect.

The intuition behind this identification result is that under the assumptions in graph 12, one can actually identify the joint effect of Z_i and M_i on D_i, which is what Pearl (2001) and Acharya, Blackwell, and Sen (2016) call the "controlled direct effect". For those individuals that shift their treatment uptake as a result of this hypothetical joint intervention, the effect of D_i on Y_i is then also identified. There are additional relevance and monotonicity assumptions needed, which are very similar to the usual LATE assumptions. We discuss these in more detail in the appendix.

We summarize all of these identification results in the following proposition:

Proposition Under the assumptions in graph (a) of figure 11, the CIA

$$D_i(Z_i = z), Y_i(D_i = d) \perp\!\!\!\perp Z_i | X_i, M_i$$

holds and under the usual monotonicity and relevance assumption, the LATE estimand

$$E[Y_i(D_i = 1) - Y_i(D_i = 0) | D_i(Z_i = 1) > D_i(Z_i = 0), X_i, M_i]$$

is identified.

Under the assumptions depicted in graphs (b) of figure 11, the CIA does not hold with any conditioning set.

Under the assumptions depicted in graphs (c) of figure 11, the CIA does hold conditional on X_i, but not conditional on M_i.

Under the assumptions depicted in figure 12, the CIA

$$D_i(Z_i = z, M_i = m), Y_i(D_i = d) \perp\!\!\!\perp Z_i | X_i, M_i$$

holds. If additionally $P(D_i(Z_i = 1, M_i = m) \geqslant D_i(Z_i = 0, M_i = m) | X_i) = 1$ ("partial" monotonicity) and $E[D_i | Z_i = 1, M_i = m, X_i] - E[D_i | Z_i = 0, M_i = m, X_i] \neq 0$ (relevance) hold, the LATE estimand

$$E[Y_i(D_i = 1) - Y_i(D_i = 0) | D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m), X_i]$$

is identified.

Proof: See appendix.

3.1.5 *Judging and Testing the Causal Assumptions*

In sum, what are the implications of these results for applied researchers if they suspect that Z_i influences M_i? We emphasize that only the restrictive sets of assumptions in figure 11 (a) and figure 12 allow for IV identification by conditioning on X_i and M_i. Again, if researchers think that the instrument may influence Y_i through variables M_i, they need to rule out confounders that may affect M_i and Y_i either directly or through D_i. We also emphasize that researchers must not condition on mediators of the $D_i \to Y_i$ relationship. This causes inconsistencies even when instruments are unconditionally valid. We now return to some of the empirical applications that motivated our research and focus on the validity of various tests that were proposed to scrutinize instrumental validity in the face of variables influenced by the instrument.

In general, robustness tests rely on determining "core" and additional control variables such that 1) identification holds with core controls, but also with additional controls and 2) there must be a chance that the robustness test fails if the assumptions are incorrect (Chen and Pearl, 2015; White and Lu, 2011a). Regarding condition 1), if one knew that one of the sets is incorrect a priori, then there would be no point in testing, as one would have to stick to the other, correct set of controls anyways. The problem with many applied papers using IV and

post-instrument variables is that they violate this condition. If one allows for the fact that the IV impacts on Y_i directly over M_i, then either the instrument is completely invalid – but one can engage in a sensitivity analysis, as shown below –, or one needs to adjust for M_i.

For example, Wucherpfennig, Hunziker, and Cederman (2016) acknowledge the possibility of various post-instrument variables and try to mitigate such concerns by adjusting for these as a robustness test. They report that estimates under either adjustment set are similar. Such a strategy is also undertaken by Kern and Hainmueller (2009) and Spenkuch and Tillmann (2017). It turns out that this testing strategy is misleading. To see why, consider first graph (a) in figure 11. In this situation, M_i-adjusted IV estimation identifies a LATE, whereas unadjusted estimates will be different and will exhibit asymptotic bias. In situations like graph (b) in figure 11, Z_i is not a valid instrument *under either adjustment*, and there are no sets of observed variables that are d-separated, so there is no way to empirically test this graph. In graph (c), M_i-adjusted IV estimates will differ, just like in graph (a), but now the *unadjusted* estimators converge to a LATE, whereas the adjusted estimates are biased. Accordingly, researchers cannot circumvent to commit themselves to causal assumptions a priori in situations like these. Comparing adjusted and unadjusted estimates is, in general, misleading: Both equal and unequal estimates may come from a real-world process where the variable Z_i is a valid instrument unconditionally, conditional on M_i, or in neither case.

There is a much more sensible testing strategy for variables that researchers think are influenced by the instrument. One situation in the which causal assumptions we have proposed are sharp enough that they allow a test is graph (a) of figure 11. In this graph, D_i and M_i are connected via the $D_i \leftarrow Z \rightarrow M_i$ path, and additional blocked paths running over the collider Y_i. Accordingly, Z_i (and X_i, as usual) d-separate D_i and M_i, and these two variables should therefore be conditionally independent in the population. This can be tested by estimating $E[D_i|M_i, Z_i, X_i]$ as a function of M_i, which is simply the first-stage that is often reported by researchers. However, the focus normally rests on the partial association between the instrument Z_i and D_i (e.g., for testing whether the instrument is weak), while the test we propose rests on the partial association between the

post-instrument variable M_i and D_i. Specifically, graph (a) of figure 11 suggests that the coefficient of a linear regression of D_i on M_i, controlling for Z_i and X_i, is zero (under a correct regression specification and appropriate standard errors). This test may seem unintuitive at first glance because it does not directly check for associations between the instrument and other variables. However, it is the only test that can be justified by relatively weak assumptions. We note that tests for ignorability of the treatment using proxies of unobserved confounders take a similar indirect route (Pei, Pischke, and Schwandt, 2017; White and Chalak, 2010).[7]

What if the test fails, i.e., the independence relationship is empirically violated? In this case, at least one open path between D_i and M_i must exist, like in graphs (b) and (c) of figure 11, or as in figure 12. Accordingly, researchers should consider a priori which of these paths may exist. Again, the possibility of causal cycles must be ruled out a priori to ensure that any of the conclusions we presented are valid.

3.2 A NEW SENSITIVITY ANALYSIS

We have shown that instruments for a causal effect may not be valid when they affect other variables that are driven by unobserved confounders and also affect the outcome of interest. Specifically, conditioning on these other variables M_i oftentimes will not achieve identification. In this section, we propose a new semi-parametric sensitivity analysis for situations where the M_i variable is driven by unobserved confounders. Our approach is based on the fact that we can often assess the effect of the instrument on the M_i variable, which provides useful information to assess the bias introduced by the direct effect of the instrument. This goes beyond other recent approaches (Conley, Hansen, and Rossi, 2012; Van Kippersluis and Rietveld, 2018) that rely completely on researchers' judgments on the sign and magnitude of the direct effect of the instrument. In contrast, our approach can use sample information. Furthermore, we relax parametric assumptions (e.g., constant effects) that are of-

7 Graph (c) of 11 also has a testable implication: $Z_i \perp\!\!\!\perp M_i | D_i, X_i$. This again is a highly non-standard test (as explained, conditioning on D_i leads to misleading inferences in all other situations). In most situations, M_i will also be driven by U_i. We discussed this graph to illustrate the mechanics of conditioning on a descendant of a collider.

ten made in the literature. Our model for Y_i, D_i, M_i looks as follows:

$$Y_i = \mu_Y + \beta_i D_i + \gamma_i M_i + \lambda_{1i}^{'} X_i + \epsilon_{1i} \tag{39}$$

$$D_i = \mu_D + \alpha_i Z_i + \pi_i M_i + \lambda_{2i}^{'} X_i + \epsilon_{2i} \tag{40}$$

$$M_i = \mu_M + \delta_i Z_i + \lambda_{3i}^{'} X_i + \epsilon_{3i}. \tag{41}$$

In this model, all causal effects vary across individuals in a fairley unrestricted fashion, and so are random variables (see Imai and Yamamoto (2013) for a similar setup). X_i is a vector of controls. We assume $E[\epsilon_{1i}] = E[\epsilon_{2i}] = E[\epsilon_{3i}] = 0$ without loss of generality. Importantly, our sensitivity analysis is consistent with graphs (a) and (b) graphs in figure 11, and additionally allows for M_i to affect D_i.[8]

We make a series of further assumptions that are enumerated in the appendix. We here give an intuitive summary. The first assumption follows from graphs (a) and (b) in figure 11. It requires that there are no unblocked backdoor paths from Z_i to any of D_i, M_i, Y_i, and that there is no direct effect of Z_i on Y_i save for the effects through D_i and M_i. The second assumption states that Z_i affects D_i monotonically, which again is a standard assumption. The third assumption requires Z_i to also affect M_i monotonically. Both monotonicity assumptions restrict π_i, so that in most situations arguments for one of these to plausible also make the other plausible. However, they are logically independent (we expand on this in the appendix). Finally, we assume that the covariance of the potential outcomes $(M(0), M(1))$ is non-negative. While the third assumption on monotonicity implies that this covariance cannot be too negative (Nutz and Wang, 2020), assuming it is non-negative greatly simplifies the analysis.

8 In graph (c), a sensitivity analysis would only be necessary if Z_i affected M_i directly. However, β_i would then no longer describe the total effect of D_i, which is of primary interest in most analyses.

In the appendix, we show that under these assumptions one can bound the weighted causal effect of D_i on Y_i, $E\left[\frac{\alpha_i + \delta_i\pi_i}{E[\alpha_i + \delta_i\pi_i]}\beta_i\right]$. The relevant bias term is

$$E[\delta_i\gamma_i] = E[\delta_i]E[\gamma_i] + cov(\delta_i, \gamma_i). \quad (42)$$

Here, $E[\delta_i]$ is the average causal effect of Z on M, which can be estimated from the data. $E[\gamma_i]$ is the direct effect of M on Y, which is the first sensitivity parameter. If treatment effects were constant, it would be the only unknown. However, if treatment effects vary and unobserved confounders impact on both M and Y, the individual-level effects δ_i and γ_i will be correlated, and the covariance term will be different from zero (Glynn, 2012). In our Vietnam draft running example, if unobserved parental SES influences the decision to attend college (M) as well as later wages (Y), then for men with low parental SES, the effect of the draft on choosing college will be relatively large (because they are more likely to be at the margin when it comes to deciding for or against college, Card (1999)). And we would expect their effect of college on earnings also to be relatively large because they have higher potential to benefit (Brand and Xie, 2010). Accordingly, the covariance would be positive. Taken together, this could lead to large bias, even if the constituent average causal effects are small. It is this potential for bias that previous approaches to sensitivity analysis have neglected (Conley, Hansen, and Rossi, 2012; Van Kippersluis and Rietveld, 2018).

We show in the appendix that one can use that data to bound this covariance term. Intuitively, the bounds are larger, the larger the standard deviation of M and the effect of Z on its standard deviation is. The second sensitivity parameter then is the standard deviation of γ_i, σ_{γ_i}. This quantity is in the same units as $E[\gamma_i]$, and describes how much γ_i typically varies. As illustrated, this standard deviation may be quite large even if mean effects are thought to be small.

Finally, we can extend the sensitivity analysis to situations where the post-instrument variable M may be measured with error. This is of special interest because often researchers are made aware of potential violations of the exclusion restriction after initial data collection. Although they then might gather some measure of a candidate M_i variable, it may well be af-

fected by measurement error. It turns out that such an error-ridden measure is still informative and can be used for sensitivity analysis.

We formalize this by complementing the model in equations 39 - 41 with a model for $M_i^\star$, the observed measure of the now unobserved M_i:

$$M_i^\star = M_i + \eta_i \tag{43}$$

and by assuming $Z_i, M_i \perp\!\!\!\perp \eta_i$ and $E[\eta_i] = 0$. This is "classical" measurement error. In the appendix, we show that the resulting estimator for the bounds stays the same, although measurement error does indeed widen the bounds compared to a situation without measurement error.

3.3 AN ILLUSTRATION OF THE PROPOSED METHODOLOGY

We illustrate our testing approach and the new sensitivity analysis using data from Spenkuch and Tillmann (2017). One aim of this paper is to estimate the effect of Catholicism on the vote share of the national socialists (NSDAP) in Germany in 1932. The data used is on the county-level and comprises official election results and census data on the share of Catholics, protestants, and other religions, as well as extensive socio-economic information like unemployment rates in various demographic subgroups. Since the authors cannot rule out unobserved confounders between religious composition and the Nazi vote share, they suggest using a county's official religion measured in 1624 as an instrument for the effect of religion on the propensity to vote for the NSDAP. They discuss evidence that the historical county denomination was largely idiosyncratic, with the exception of a few observable factors, for which they adjust in their statistical analysis. In our framework, these variables correspond to pre-instrument covariates X_i.

Spenkuch and Tillmann (2017, p. 9) then further assert that in order for this historical variable to be a valid instrument for the effect of interest, "it may influence voters' decisions to support the NSDAP only through its impact on covariates that are included in the regression". We take this as indication that past religious composition Z_i may have affected, for example, the economic situation in counties in the 1930s, which we

conceptualize as M_i variables. One well-known mechanism for such an effect is Max Weber's hypothesis of a "protestant work ethic". Furthermore, it is plausible that such economic variables also exerted a strong influence on the Nazi vote share. Accordingly, the instrument would be valid if we faced the situation of graph (a) in figure 11. Assuming the IV is valid, Spenkuch and Tillmann (2017) estimate that a one percentage point increase in the share of Catholics in a county decreased the NSDAP votes share by about 0.27 percentage points. The estimate is quite precise (standard error of about 0.03) and very large. It is substantively and statistically indistinguishable from OLS estimates, where Catholicism explains about 40% of the variance in NSDAP votes.

From this alone, it is clear that only strong deviations from the IV assumption can change the substantive conclusions. We concentrate on one single M_i variable measuring a highly relevant economic fundamental: The county-level unemployment rate among blue-collar workers. We first employ our diagnostic test and check whether this variable is independent of the treatment, conditional on the instrument and other controls, using the most extensive specification employed in Spenkuch and Tillmann (2017). We find a significant dependency, with a coefficient of .45 and an estimated standard error of .18 ($p < 0.02$). This leads us to reject a scenario like graph (a) in figure 11. It is conceivable that unobserved confounders do not only influence religious composition and NSDAP vote share, but also unemployment rates, so that graph (b) is more realistic. However, in this case, Z_i (the historical denomination of a county) is not a valid instrument anymore, neither unconditionally nor conditional on M_i. Accordingly, we proceed with our sensitivity analysis.

The next step is to gather some evidence on the mean and variance effects of the instrument on M_i. The mean effect is very small and indistinguishable from zero (95% CI: [-0.03, 0.03]). The effect of Z_i on the standard deviation of M_i, however, is quite large (95% CI: [2.45, 4.26]). This leads us to reject that causal effects are constant. Furthermore, it is possible that the covariance between causal effects is also large. In sum, we do not perform the sensitivity analysis with respect to the mean parameters, as the "first-stage" already is very small, and the

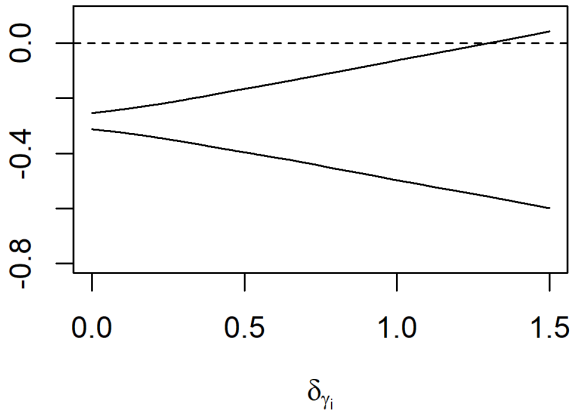

Figure 13: 95% confidence intervals for the effect of Catholicism on NSDAP vote shares, as a function of the variability of the effect of unemployment on NSDAP vote shares. Based on data from Spenkuch and Tillmann (2017).

results will be robust. Instead, we concentrate on sensitivity to causal effect heterogeneity.

Figure 13 plots the 95% confidence interval for the causal effect of Catholicism on NSDAP vote, as a function of the variability of the effect of the unemployment rate on NSDAP votes shares. Only with a standard deviation beyond ca.1.25 does the effect becomes statistically insignificant. Is this reasonable to expect? Due to severe data limitations, the empirical literature on Weimar elections focusses on descriptive inferences (King et al., 2008), so that it cannot directly inform our assessment on the magnitude of σ_{γ_i}. Spenkuch and Tillmann (2017)'s own estimates for mean effects of unemployment, which they do not claim are causal effects, are negative and at most as large as the effect of Catholicism. The contemporary literature on the causes of extreme right voting (Arzheimer, 2009; Jackman and Volpert, 1996) finds positive effects of both individual unemployment and aggregate unemployment rates. This suggests some variability in effects. However, even if the effect of unemployment rates would vary uniformly between, say, -0.75 and 0.5 percentage point increases across counties, the implied standard deviation would only be about 0.36. While this introduces some additional uncertainty, the main inference is robust.

3.4 CONCLUSION

Many applied researchers use instrumental variables in settings where they try to "control away" a direct effect of the instrument on the outcome by measuring other variables M. In this paper, we explained why this strategy only works under specific, restrictive assumptions. Using causal graphs, we highlighted the asymmetric role of pre- and post-instrument covariates: While adjustment for the former is often necessary and unproblematic, statistical control for the latter has to be taken with extreme caution. We showed that with direct effects of the instrument through M, some local average treatment effects may be identified, but we also highlighted various sources of asymptotic bias. We discussed the limited value of existing robustness tests and provided a more suitable test of a specific set of identification assumptions. Finally, we introduced a sensitivity analysis as an alternative, and illustrated it using the IV analysis of Spenkuch and Tillmann (2017). Here, it became clear that both mean effects as well as the variability of causal effect may play an important role for the sensitivity of estimates, although the inference of the paper appears to be very robust.

We conclude by providing a checklist for applied researchers that want to utilize a (potential) instrumental variable that may have a direct effect on the outcome through another variable:

1. Based on substantive knowledge, determine which of the graphs discussed in this paper seems plausible for your research design. Specifically, be clear about which variables are confounders X_i that influence Z_i, D_i, and Y_i. and which variables M_i are driven by Z_i or D_i.

2. If M_i is a mediator and not directly driven by Z_i (as in graph (c) of figure 11), proceed with standard estimation routines like 2SLS, where you condition only on X_i.

3. If your assumptions are equivalent to those in graph (a) of figure 11, implement the diagnostic test by checking whether D_i and M_i are independent conditional on Z_i. If they are, condition on X_i and M_i in your statistical analysis.

4. If the test fails, reconsider your assumptions. Only the assumptions in figure 12 allow for conditional dependency

between D_i and M_i and identification based on adjustment for X_i and M_i.

5. If prior knowledge or the diagnostic test leads to the conclusion that Z_i directly influences M_i and that the unobserved confounder also influences M_i (as in graph (b) of figure 11), identification is not possible. Perform estimation conditional only on X_i and then use our sensitivity analysis to assess whether substantive conclusions still hold.

APPENDIX

3.A PROOF OF THE PROPOSITION

For ease of exposition, we first introduce some useful properties of conditional independence:

Lemma. *(Dawid, 1979) If* $X_i \perp\!\!\!\perp Y_i | Z_i$ *and* U_i *is a function of* X_i*, then 1)* $U_i \perp\!\!\!\perp Y_i | Z_i$ *and 2)* $X_i \perp\!\!\!\perp Y_i | Z_i, U_i$.

Lemma. *(Contraction, Pearl (2009))* $X_i \perp\!\!\!\perp Y_i | Z_i$ *and* $X_i \perp\!\!\!\perp W_i | Z_i, Y_i$ *imply* $X_i \perp\!\!\!\perp Y_i, W_i | Z_i$.

Lemma. $Z_i \perp\!\!\!\perp U_i | X_i$ *implies* $Z_i \perp\!\!\!\perp f(U_i), g(U_i) | X_i$*, where* f, g *are arbitrary functions.*

Proof. $Z_i \perp\!\!\!\perp U_i | X_i$ implies $Z_i \perp\!\!\!\perp f(U_i) | X_i$ as well as $Z_i \perp\!\!\!\perp U_i | X_i, f(U_i)$ by lemma 1. The latter then similarly implies $Z_i \perp\!\!\!\perp g(U_i) | X_i, f(U_i)$. By contraction, we then have $Z_i \perp\!\!\!\perp f(U_i), g(U_i) | X_i$. □

We can now prove the statements in the main text. Throughout, we will assume there are additional observed confounders X_i influencing all observed variables.

Proof of the Proposition. In graph (a) of figure 11, we have $Y_i(D_i = d) = f_y(d, M_i, X_i, U_i)$ and $D_i(Z_i = z) = f_d(z, X_i, U_i)$. By d-separation, the graph implies $Z_i \perp\!\!\!\perp U_i | X_i, M_i$. By Lemma 3, this implies $Y_i(D_i = d), D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i, M_i$. Identification of the X_i, M_i-specific LATE then follows as in Angrist, Imbens, and Rubin (1996).

In graph (b) of figure 11, $Y_i(D_i = d) = f_y(d, M_i, X_i, U_i) = f_y(d, f_m(Z_i, X_i, U_i), X_i, U_i)$, which depends on Z_i. Conditioning

on X_i does not block this dependency. Conditioning on X_i, M_i makes Z_i and U_i dependent, so the CIA is generally violated. However, $D_i(Z_i = z) = f_d(z, X_i, U_i)$, and $D_i \perp\!\!\!\perp U_i | X_i$ by d-separation, so $Z_i \perp\!\!\!\perp D_i(Z_i = z) | X_i$ holds and the ATE of Z_i on D_i is identified.

In graph (c) of figure 11, $Y_i(D_i = d) = f_y(d, X_i, U_i)$ and $D_i(Z_i = z) = f_d(z, X_i, U_i)$. d-separation implies $Z_i \perp\!\!\!\perp U_i | X_i$, so by lemma 3, $Y_i(D_i = d), D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i$. However, conditioning on M_i makes Z_i and U_i dependent.

In figure 12, we have

$$Y_i(D_i = d), D_i(Z_i = z, M_i = m) \perp\!\!\!\perp Z_i | X_i, M_i$$

(CIA.2)

First, in this graph, $Y_i(D_i = d) = f_y(d, M_i, X_i, U_i)$ and $D_i(Z_i = z, M_i = m) = f_d(z, m, X_i, U_i)$. By d-separation, we have $Z_i \perp\!\!\!\perp U_i | X_i, M_i$. Lemma 3 then implies CIA.2. Additionally, we assume

$P(D_i(Z_i = 1, M_i = m) \geqslant D_i(Z_i = 0, M_i = m) | X_i = x) = 1$ for all m, x (partial monotonicity)

$E[D_i | Z_i = 1, M_i = m] - E[D_i | Z_i = 0, M_i = m] \neq 0$ for all m, x (relevance)

Consider the X_i, M_i-adjusted Wald estimator

$$\frac{E[Y_i | Z_i = 1, M_i = m, X_i] - E[Y_i | Z_i = 0, M_i = m, X_i]}{E[D_i | Z_i = 1, M_i = m] - E[D_i | Z = 0, M_i = m]}$$

Under the above assumptions, the numerator evaluates to

$$E[Y_i | Z_i = 1, M_i = m, X_i] - E[Y_i | Z_i = 0, M_i = m, X_i] =$$

$$E[(Y_i(D = 1) - Y_i(D = 0))(D_i(Z_i = 1, M_i = m) - D_i(Z_i = 0, M_i = m)) | M_i = m] =$$

$$E[Y_i(D = 1) - Y_i(D = 0) | D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m)] \times$$

$$P(D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m) | M_i = m)$$

The first step follows from consistency (see Pearl 2009, Corollary 7.3.2) and CIA.2, and the second step follows from partial monotonicity. The denominator is

$$E[D_i(Z_i = 1, M_i = m) | M_i = m] - E[D_i(Z_i = 0, M_i = m) | M_i = m] =$$

$$P(D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m)|M_i = m)$$

The first step follows from consistency and CIA.2, and the second step follows from partial monotonicity. Accordingly, the Wald estimator evaluates to

$$E[Y_i(D = 1) - Y_i(D = 0)|D_i(Y_i = 1, M_i = m) > D_i(Y_i = 0, M_i = m)]$$

□

To assess the relationship between the traditional monotonicity assumption and partial monotonicity, consider the case of binary Z_i and binary M_i, and no covariates. In this case, a saturated structural model for D_i without any functional-form assumptions can be written

$$D_i = \alpha_i + \beta_{i1}Z_i + \beta_{i2}M_i + \beta_{i3}Z_iM_i$$

where $\alpha_i = E[D_i(Z_i = 0, M_i = 0)]$, $\beta_{1i} = D_i(Z_i = 1, M_i = 0) - D_i(Z_i = 0, M_i = 0)$, etc. Monotonicity requires $D_i(Z_i = 1) \geqslant D_i(Z_i = 0)$ for all i, which is equivalent to stating that $\beta_{1i} + \beta_{3i}M_i \geqslant 0$ for all i. This restricts the joint distribution of $(\beta_{1i}, \beta_{3i}, M_i)$, and means that for no individual, effects and choices of M_i are correlated such that increases in the instrument induce decreases in the treatment. Partial monotonicity is equivalent to the requirement that $\beta_{1i} + \beta_{3i}m \geqslant 0$ for all m and i, where m is constant. This restricts the distribution of (β_{1i}, β_{3i}). In theory, there could be fine-tuned distributions of $\beta_{1i}, \beta_{3i}, M_i$ where monotonicity holds but partial monotonicity does not (e.g., $\beta_{1i} > 0$ for all i, $\beta_{3i} = 0$ for those with $M_i = 1$, but $\beta_{3i} < 0$ and $|\beta_{3i}| > \beta_{1i}$ for those with $M_i = 0$). However, it seems natural to assume that the restrictions on β_{1i}, β_{3i} also hold when suitable restrictions on $\beta_{1i}, \beta_{3i}, M_i$ are plausible.

3.B DERIVATION OF THE SENSITIVITY ANALYSIS

The structural model suggests estimation of all regression functions using linear models where the control variables X_i enter separately. Therefore, we leave the conditioning on X_i implicit in the following; all variables can be thought of as having partialled out their correlation with X_i. Consistent with this, we also assume that our sensitivity parameters are independent of

X_i (see Knox, Lowe, and Mummolo (2020, p. 11) for a similar approach).

Formally, our assumptions in addition to 39–41 are

$$Z_i \perp\!\!\!\perp (\beta_i, \gamma_i, \alpha_i, \pi_i, \delta_i, \epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i}) \tag{44}$$

$$P(\alpha_i + \delta_i \pi_i \geqslant 0) = 1 \tag{45}$$

$$P(\delta_i \geqslant 0) = 1 \tag{46}$$

$$cov(M_i(0), M_i(1)) \geqslant 0 \tag{47}$$

Under these assumptions, we have

$$\begin{aligned} &E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \\ &\quad E[\beta_i(\alpha_i + \delta_i \pi_i)] + E[\delta_i \gamma_i]. \end{aligned} \tag{48}$$

This holds because Z_i is independent of all causal effects and the error terms. $E[\delta_i \gamma_i]$ is the bias term we need to bound.

Using similar reasoning, we also have

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 1] = E[\alpha_i + \delta_i \pi_i] \tag{49}$$

and

$$E[M_i|Z_i = 1] - E[M_i|Z_i = 0] = E[\delta_i]. \tag{50}$$

3.B.1 *With measured* M_i

Rewrite the bias term as

$$E[\delta_i \gamma_i] = cov(\delta_i, \gamma_i) + E[\delta_i]E[\gamma_i]. \tag{51}$$

In the second term, $E[\delta_i]$ is point-identified as $E[M_i|Z_i = 1] - E[M_i|Z_i = 0]$, while $E[\gamma_i]$ will be a sensitivity parameter.

Further rewrite

$$cov(\delta_i, \gamma_i) = cor(\delta_i, \gamma_i)\sigma_{\delta_i}\sigma_{\gamma_i}. \tag{52}$$

In this latter term, we can decompose σ_{δ_i} as

$$\sqrt{var(M_i(1)) + var(M_i(0)) - 2cov(M_i(1), M_i(0))}. \tag{53}$$

The variance terms are nonparametrically point-identified as $var(M_i|Z_i = z)$. Regarding the covariance, intuition might suggest that monotonicity ($M_i(1) \geqslant M_i(0)$) implies that it is positive, but one can create joint distributions of $(M_i(1), M_i(0))$ where this is not the case. However, the Frechét-Hoeffding bounds (e.g. Aronow, Green, Lee, et al. (2014)) for this quantity using the marginals are not sharp, because the monotonicity does, in fact, improve the lower bound. Very recent work characterizes this lower bound under monotonicity (Nutz and Wang, 2020). Since we are not aware of research on how to estimate this bound, especially with covariates, we make the simplifying assumption that $cov(M_i(1), M_i(0)) \geqslant 0$. Using this, an upper bound for equation 53 is

$$\sqrt{var(M_i|Z_i = 1) + var(M_i|Z_i = 0)}. \tag{54}$$

Further using $-1 \leqslant cor(\delta_i, \gamma_i) \leqslant 1$, we can bound equation 52 as

$$\begin{aligned} -\sqrt{(var(M_i|Z_i = 1) + var(M_i|Z_i = 0))}\sigma_{\gamma_i} \\ \leqslant cov(\delta_i, \gamma_i) \leqslant \\ \sqrt{(var(M_i|Z_i = 1) + var(M_i|Z_i = 0))}\sigma_{\gamma_i}, \end{aligned} \tag{55}$$

where σ_{γ_i}, the standard deviation of the direct causal effect of M_i on Y_i, is the second sensitivity parameter.

Collecting terms and rearranging, we have

$$\begin{aligned} &\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} - \frac{1}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \times \\ &\{(E[M_i|Z_i = 1] - E[M_i|Z_i = 0])E[\gamma_i] + \sqrt{var(M|Z = 1) + var(M|Z = 0)}\sigma_{\gamma_i}\} \\ &\leqslant E\left[\frac{\alpha_i + \delta_i\pi_i}{E[\alpha_i + \delta_i\pi_i]}\beta_i\right] \leqslant \\ &\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} - \frac{1}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \times \\ &\{(E[M_i|Z_i = 1] - E[M_i|Z_i = 0])E[\gamma_i] - \sqrt{var(M|Z = 1) + var(M|Z = 0)}\sigma_{\gamma_i}\}, \end{aligned} \tag{56}$$

if $\frac{E[M_i|Z_i = 1] - E[M_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$ is positive. If it is negative, the inequality signs reverse.

3.B.2 *With mismeasured* M_i

As before, we want to gain information on the bias term (equation 51) from the data. $E[\delta_i]$ remains identified under the measurement model in equation 43 and the stated assumptions on the measurement error: $E[M_i^\star|Z = 1] - E[M_i^\star|Z = 0] = E[M_i + \eta_i|Z = 1] - E[M_i + \eta_i|Z = 0] = E[M_i|Z = 1] - E[M_i|Z = 0] = E[\delta_i]$.

It further turns out that the variances $var(M_i(z))$ are not point-identified anymore, although they can be bounded from above by the same quantities as in the case without measurement error. Accordingly, the resulting bounds for the sensitivity analysis do not change. To see why, consider

$$\begin{aligned} var(M_i(z)) = var(M_i|Z_i = z) = var(M_i^\star - \eta_i|Z = z) = \\ var(M_i^\star|Z = z) + var(\eta_i|Z = z) - 2cov(M_i^\star, \eta_i|Z = z) = \\ var(M_i^\star|Z = z) + var(\eta_i) - 2cov(M_i^\star, \eta_i|Z = z). \end{aligned} \tag{57}$$

Regarding this last term, we have

$$\begin{aligned} cov(M_i^\star, \eta_i|Z = z) = cov(M_i + \eta_i, \eta_i|Z = z) = \\ cov(M_i, \eta_i|Z = z) + var(\eta_i|Z = z) = var(\eta_i). \end{aligned} \tag{58}$$

Accordingly,

$$var(M_i(z)) = var(M_i^\star|Z = z) - var(\eta_i) \leqslant var(M_i^\star|Z = z). \tag{59}$$

This bound could be improved upon if we could improve the trivial zero lower bound for $var(\eta_i)$. However, it is only possible to bound $var(\eta_i)$ from above using $var(M_i)$.

In sum, equation 59 shows that the observed conditional variance of the measurement is equal to or larger than the marginal variance of the potential outcome of the actual M_i variable. If measurement error is large, the empirical estimate will be far away from zero, even though the true marginal variance might

be close or equal to zero. This is the information loss incurred by the measurement error.

Accordingly, the bounds in equation 56 remain valid, substituting $M_i^\star$ for M_i.

3.B.3 *Implementation*

For implementing the sensitivity analysis, we need to make a number of choices for estimation and inference. As stated before, and consistent with most IV applications, estimation of the mean differences in equation 56 can be pursued using two-stage least squares. For the variance terms, we pick corresponding linear conditional variance models. We first estimate auxiliary mean regressions, and then use linear models for the squared residuals Shalizi, 2019, p. 217. We estimate $\mathrm{var}(M|Z = z)$ economically by mean-centering all controls, such that our estimate for $\mathrm{var}(M|Z = 1) + \mathrm{var}(M|Z = 0)$ is $2a + b$, where a is the constant from the variance regression, and b is the coefficient on Z.

Finally, we use the bootstrap to estimate the sampling distribution of the resulting estimator.

4

GRAPHICAL CAUSAL MODELS FOR SURVEY INFERENCE

Co-authored with Peter Selb. See Part iv for author's contribution.

Surveys figure among the most prominent data collection tools for social research. Regardless of whether one is dealing with an election poll, health survey, or census, the general purpose of a survey is to provide information about the distribution of some individual attribute (e.g., an attitude, behavior, or social characteristic) in a population, usually based on self-reports in a sample of individuals selected from the population. The inferences involved – from self-reports to underlying attributes (i.e., measurement) and from samples to target populations (i.e., generalization) – are prone to various errors which may arise in the design, collection, and analysis of survey data. While the Total Survey Error (TSE) framework has emerged as the dominant conceptual foundation for studying the sources of error and ways of assessing and improving survey data quality (e.g., Groves et al., 2009), some influential work in survey methodology has used graphical models to illustrate errors (e.g., Groves, 2006; Groves and Peytcheva, 2008; Kreuter and Olson, 2011; Olson, 2019; Schafer and Graham, 2002). However, their usage has largely remained informal.

In this article, we demonstrate how the utility of such models for survey research can be expanded by utilizing rules and insights from the literature on Directed Acyclic Graphs (e.g., Morgan and Winship, 2015; Pearl, 2009). In doing so, we provide survey researchers with a formal yet intuitive tool to encode and communicate causal assumptions about the collection of survey data and to support the choice of suitable adjustment strategies. We thus contribute to a development Groves and Lyberg (2010, p. 866) urged in their critical appraisal of the state of survey methodology a decade ago:

> Missing in the history of the total survey error formulation is the partnership between scientists who study the causes of the behavior producing the statis-

tical error and the statistical models used to describe them.

After elaborating on the tight connection between graphs and statistical dependencies, we show how classic assumptions made in the adjustment for survey nonresponse ("missing at random", etc.) can be intuitively yet rigorously formalized using graphs. Inter alia, we emphasize the importance of using causal instead of correlational language and discuss how "collider bias" can explain findings of vote validation studies. We then showcase the potential of causal graphs by analyzing multiple selection stages. Here, we closely follow the TSE framework and go beyond existing analyses of sample selection using DAGs (e.g., Bareinboim, Tian, and Pearl, 2014; Daniel et al., 2012; Moreno-Betancur et al., 2018; Thoemmes and Mohan, 2015). Specifically, we discuss the implicit strong assumptions behind the widespread practice of using "design weights" (and only those) when analyzing survey data. Next, we show that requirements formulated in the literature (Kreuter et al., 2010; Peytchev, Presser, and Zhang, 2018; Sakshaug and Antoni, 2019) for valid adjustment variables to be correlated with response indicators and survey outcomes to be ill-justified.

We conclude our paper by outlining future areas of research where causal graphs can be used to better understand the assumptions of survey methods or to develop new methods: regression adjustment, sample selection models based on instrumental variables, "transporting" causal effect estimates using observational data, sensitivity analysis, measurement error, and efficient covariate control.

4.1 PROBABILITY BASICS

In the following sections, we introduce some basic rules and principles from causal graph and probability theory, which are necessary to follow the subsequent discussion. We begin with *random variables*, Y and X, and their marginal population distributions $P(Y)$ and $P(X)$. In our running example, $P(Y)$ is the distribution of candidate preferences in the population, and $P(X)$ the distribution of educational degrees.

$P(Y|X)$ describes the conditional probability of Y given X. If this probability actually varies as a function of X, we say that X and Y are *dependent*, which we will often loosely refer to as

correlated. If $P(Y|X) = P(Y)$, so that $P(Y)$ is constant across values of X, we say that Y and X are *independent* or uncorrelated. $E[Y]$ denotes the population mean of Y, and $E[Y|X]$ denotes the conditional mean of Y given X. The *law of total probability* asserts that

$$P(Y) = \sum_x P(Y|X = x)P(X = x). \tag{60}$$

That is, one can always divide the distribution of a variable Y into distributions of Y conditional on $X = x$ and then get back to the marginal distribution $P(Y)$ by summing over and weighting the conditional distributions with their relative size $P(X = x)$ provided that $P(X = x) > 0$. This works for means as well (*law of iterated expectations*):

$$E[Y] = \sum_x [E|X = x]P(X = x). \tag{61}$$

The law of total probability is the basis for post-stratification and other survey adjustment strategies (e.g., Lohr, 2019). In such a case, one knows the group (x-specific) distribution of Y from the survey. Then one can estimate the overall population distribution of Y by weighting these group distributions by the size of the groups in the population known from external sources.

4.2 GRAPH BASICS

In general terms, causal graphs encode a researcher's assumptions about the causal process that generated the data in a population of interest. They consist of variables, or *nodes*, which are linked by *arcs*. We use arrows for arcs to depict the hypothesis that one node influences the other. The presence of an arc between two nodes merely indicates that a certain causal effect might be present. On the other hand, the absence of an arc between two nodes indicates a critical assumption, namely that one does not affect the other. A *path* is a sequence of arcs that links one node to another, regardless of the direction of arrows. Retracing arcs or going through the same node twice is not allowed. A *directed* or *causal path* is traced out along arrows tail-to-head. If there is a directed path from one node to another, the former is said to be an *ancestor* of the latter, the latter a *descendant* of the former. A Directed Acyclic Graph, or DAG, contains

only directed arrows and no feedback loops (i.e., no variable is its own ancestor or descendant).

The causal assumptions encoded in a DAG constitute the theoretical model on which *identification analysis* rests. Identification analysis is concerned with estimating parameters of interest, regardless of random variability in the data due to small samples or measurement error. In this sense, it is closely related to the statistical notion of *consistency*. The main reason to focus on identification as a first step is that if a parameter cannot be identified using "infinite" and error-free data, we certainly cannot learn anything about it using finite data.[1] So, for the time being, we assume large samples and error-free measurements. We return to issues of estimation and measurement in the conclusion. Finally, the graphical literature on sample selection and missing data often uses the term *recovery* or *recoverability* of some quantity (Mohan and Pearl, 2019). We will use the terms identification, recovery, and recoverability interchangeably.

4.3 D-SEPARATION

Having discussed the basic semantics of causal graphs, we now turn to the only graphical rule on which we rely: *d-separation*. It allows us to logically infer the absence of correlations from the structure of the graph. We give here a rather dense discussion of the formalities, which will later be illustrated through various examples.

Figure 14 depicts the three basic patterns one needs to understand when determining d-separation. In the left panel, Z is a common cause, or *confounder*, of X and Y. In this scenario, Z induces a statistical association between X and Y, although X and Y do not cause each other, which is indicated by the absence of an arrow between them. In the intermediate graph, the causal effect of X on Y is *mediated* through Z. This also has the consequence of producing a correlation between X and Y because the former influences the latter. In short, we say that these two paths are *open*. Finally, in the right panel, Z is a common effect, or *collider*, on the path between X and Y. In contrast to the two

1 Often, however, parameters can be set-identified (bounded) under weak assumptions. Our interest in this article is on point identification. For bounding parameters under sample selection, see Manski (2009).

other cases, this does not produce a correlation between X and Y. Here, we say that the path is *blocked* or d-separated by the variable Z.

These patterns of association reverse once we look at the distribution of X and Y conditional on Z, i.e., $P(Y|X,Z)$. For example, this could be done by regressing Y on X while controlling for Z, or by stratification or matching on Z (e.g., Zhang, 2000). In the left graph, Z is the only common cause of X and Y. Accordingly, for units with the same value of Z, the value of X is not informative about Y, and vice versa. In the intermediate graph, conditioning on Z also blocks the information flow from X to Y. X and Y are said to be d-separated conditional on Z. However, in the collider graph on the right, an association emerges. To understand why an example is helpful.

Consider two independent binary variables X and Y and a random variable Z that is the sum of X and Y. Therefore, Z can take on the values $\{0, 1, 2\}$. X and Y may be random coin flips, so knowing the value of X does not help in predicting Y. However, conditioning on Z means that we are told its value. Knowing that Z is 1, for example, and that X is 0, we know that Y has to be 1. Conditional on Z, X and Y are dependent or d-connected. Put differently, knowing the result of a process and the value of one of its independent inputs also lets us predict the value of the other input.[2]

In sum, in Figure 14, conditioning on the intermediate variable Z blocks the path between X and Y in the first two graphs but opens the path in the right graph. For deriving whether variables X and Y are (conditionally) independent in more complex DAGs, it turns out that one can just enumerate all paths between these variables. If all of these paths are blocked, perhaps conditional on other variables Z, then we say that X and Y are d-separated (conditional on Z). If all of the variables involved are measured, this statement is testable: The distribution of Y (for some value of Z) should not change for different values of X. For instance, if one commits to a specific regression model, the test involves regressing Y on X and Z; the coefficient on X should be zero (One could also use X as the dependent and Y as the independent variable). The only reason that the coeffi-

2 The same mechanics apply if we happen to know the realization of a descendant of Z. For example, let D be a variable that takes on the value 1 when Z equals 1, otherwise 0 (so that it is a binary proxy for Z). Knowing that D equals 1 and that X equals 0 also leads to the prediction that Y equals 1.

cients can be different from zero is that the DAG is incorrectly specified and there is at least one open path.[3]

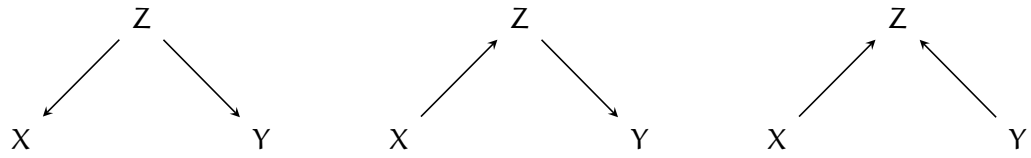

Figure 14: Basic causal patterns and d-separation: Z is a *confounder* (left), a *mediator* (center), or a *collider* (right) on the path between X and Y. In the latter scenario, the path is naturally blocked by Z, which entails that there is no statistical association between X and Y. Otherwise, paths are open and induce statistical association between X and Y.

4.4 DAGS IN CAUSAL INFERENCE

To give an idea of the role of DAGs in causal inference, we note that most empirical research that uses regression modeling to estimate causal effects explicitly or implicitly relies on the "exogeneity" of the independent variables of interest. If one employs causal graphs to visualize assumptions, a simple graphical criterion to determine the exogeneity of a variable X with respect to an outcome Y is the *backdoor criterion* (Pearl, 2009, p. 79). It requires that we look for control, or "adjustment", variables Z such that (1) no such variable is a descendant of X, and (2) Z blocks all paths between X and Y that contain an arrow into X, that is, *backdoor* paths. If such variables Z exist and can be measured, adjustment for them – be it through regression, matching, weighting, or another approach – allows estimating the causal effect of X on Y.

An extensive discussion of this issue is beyond the scope of this paper. We refer interested readers to Elwert and Winship (2014a), who concentrate on problems that occur when condition (1) of the backdoor criterion is violated. We will, however, touch upon a few themes that have already emerged in this very short discussion of causal inference. First, the DAG literature typically produces graphical criteria that are easy to check given a graph and cover many, perhaps even all, possible strategies to solve a certain problem. We will see more examples of this below. Second, such criteria fundamentally rely on the logic

3 In finite samples, the estimate will never be exactly zero.

of blocking paths, often complemented with some additional requirements (e.g., that Z may not contain descendants of X). Third, note that the backdoor criterion asks us to consider the relationship (i.e., backdoor paths) between X and Y in the graph. While this may seem like a trivial requirement, it implies that such statements as "X is exogenous" are not precise; X can only be exogenous with respect to an outcome Y, and in fact, with respect to a causal model.[4] A similar requirement also occurs for inferences from selective samples.

4.5 SURVEY INFERENCE FROM A DAG PERSPECTIVE

In this section, we provide detailed explanations of the causal structures behind nonresponse biases and show how existing assumptions for adjustment can be better understood using graphs. While social science scholars interested in causal effects often use survey data, the prime interest of survey researchers typically lies in the population distribution of some variable Y, perhaps given another variable X. A highly visible example which we will be referring to throughout the following exposition is election polling, where Y is a candidate or party preference, and X is a sociodemographic variable such as education. The context of elections provides ample opportunities for validating survey statistics (e.g., Ansolabehere and Hersh, 2012; Shirani-Mehr et al., 2018), and failures of polls to predict high-profile elections often trigger investigations which yield relevant insights into survey errors.

For instance, Kennedy et al. (2018) evaluate several theories as to why many polls preceding the 2016 U.S. presidential election, particularly in the Midwest, underestimated support for Trump. One theory maintains that preferences for the Democratic candidate (Hillary Clinton) were increasing in formal education in 2016, while they had been U-shaped historically. Since highly educated voters are often overrepresented in surveys, they used to "proxy" for citizens of lower education, but this ceased to be the case in 2016 when preferences differed between those two groups. Accordingly, surveys that did not adjust for education X overestimated Clinton's vote share Y.

Conceptually, selection into the survey sample can be formalized using a binary variable S that is 1 whenever an element of

4 See Pearl (2009, pp. 165–170) for a discussion of different concepts of exogeneity.

the population is included in a survey and 0 whenever it fails to be. The data one actually measures in the survey are conditional on $S = 1$; i.e., one does not observe $P(Y)$, but merely $P(Y|S = 1)$. Accordingly, sample selection is a problem because one is *forced* to condition on the variable S. From this perspective, unit nonresponse in a probability sample is equivalent to selection into a nonprobability sample, although the resulting biases are often larger and harder to adjust for in the latter case (Cornesse et al., 2020).

In graphs, such – inadvertent – conditioning is indicated by putting boxes around the selection nodes in Figure 15. In the following, we use these three simple graphs to explain how to represent the prototypical scenarios discussed in the missing data literature (e.g., Little and Rubin, 2019; Rubin, 1976): Missingness Completely At Random (MCAR), Missingness (MAR), and Missingness Not At Random (MNAR) (also see Thoemmes and Mohan, 2015).

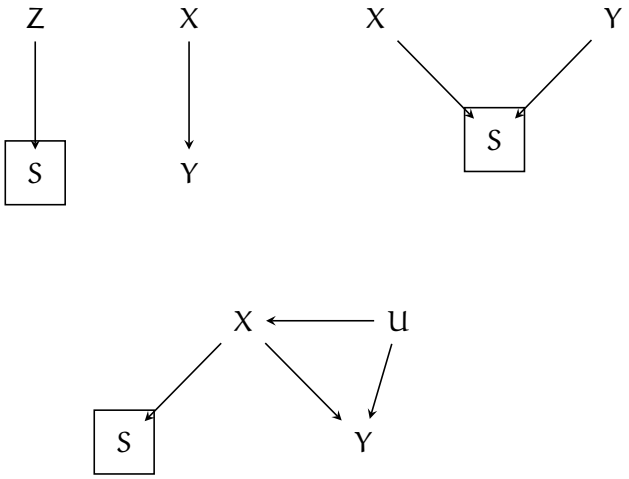

Figure 15: Prototypical selection scenarios. Left: Y is missing completely at random (MCAR), i.e., S and outcome Y have separate causes. Center: Y influences missingness so that it is missing not at random (MNAR). Right: Y is missing at random (MAR) conditional on X, because X drives S and Y.

Missingness Completely at Random (MCAR)

Suppose we are interested in estimating the population distribution of some variable Y from a sample of individuals with $S = 1$. In the left panel of Figure 15, selection node S and survey variable Y have separate sets of causes, Z and X. Since there

is no path whatsoever between S and Y, the two variables are d-separated, and we have $Y \perp\!\!\!\perp S$ (read: Y is independent of S). Accordingly, we can write:

$$P(Y|S = 1) = P(Y), \tag{62}$$

so that the distribution we observe in the sample, $P(Y|S = 1)$, asymptotically equals the population distribution we are interested in, $P(Y)$. In the terminology of the missing data literature, the selection process is *ignorable*, and Y is *missing completely at random*.

In a prototypical survey without nonresponse, Z would be the output of some randomization device and the only cause of S. This, in expectation, warrants that the selection of respondents is independent of the survey variable of interest (and, in fact, any other variables including X). Just like randomized treatment assignment in experiments, random selection of respondents is a convenient *procedural* justification of independence. Yet, random selection is not necessary for $Y \perp\!\!\!\perp S$. For instance, assume that we are dealing with an election poll in which Y is party choice and X are policy preferences, where the selection process is random sampling subject to nonresponse due to the lack of interest in politics Z among sampled individuals (see Groves, Presser, and Dipko, 2004). In this situation, the population distribution of Y can be recovered from the survey sample despite nonresponse as long as there is no open path between political interest and policy preferences. On the other hand, even an intact probability sampling scheme will not lead to independent Y and S, if the design entails unequal inclusion probabilities (e.g., for strata) unaccounted for in the analysis, and mean values of Y differed across strata. Also, any probability sampling may produce associations between Y and S by chance (see the discussion in Valliant, Dorfman, and Royall, 2000, pp. 19–21).

Missingness Not at Random (MNAR)

In the center panel of Figure 15, both X and Y influence response in the survey. It is well known that more highly educated persons (X) are overrepresented in surveys, and it was also suspected that there was a relationship between vote preferences (Y) and survey response before the 2016 election

(Kennedy et al., 2018). The graph would be consistent with both accounts. Recall that the data one actually observes are conditional on $S = 1$. If we are interested only in the marginal distribution of candidate preferences Y, we will have a problem. If, for example, supporters of a candidate trailing in the polls have a lower propensity to participate in the survey, then people that end up in the respondent pool will be more likely to have a preference for the leading candidate (e.g., Gelman et al., 2016). Accordingly, we cannot make consistent inference on $P(Y)$ since it differs from the distribution we measure, $P(Y|S = 1)$. Whenever the outcome directly affects sample inclusion, this problem cannot be solved, at least regarding the quantities of our interest.[5]

A more peculiar implication of this graph is that if the interest is in analytic statistics, in-sample correlations between independent and outcome variables can exist even though there is no relationship between these variables in the population. The existing literature on survey nonresponse does not discuss this phenomenon (Groves, 2006; Mercer et al., 2017; Peytchev, 2013; Peytchev, Carley-Baxter, and Black, 2011). For example, if we are interested in how candidate preference Y varies as a function of education X, we encounter a collider structure where we are forced to condition on selection S. This means that education and preference are correlated in the sample, although in this stylized example no such correlation exists in the overall population. Why is this the case? Assume that both education and preference have a positive effect on the probability of responding in a survey. Then, among those that actually respond ($S = 1$) and that have relatively low education, preference Y is more likely to be 1 because it has to "make up" for the low value of X in producing $S = 1$. Accordingly, we can derive that in such a case, education and preference are negatively correlated in the sample. If we allowed for an effect of X on Y or unobserved confounders to influence these variables, the result would not qualitatively change. For example, we could have a positive correlation in the population while the correlation among sampled subjects is zero or negative. Often, this is called "collider bias".[6] It represents a potential explanation

5 See Didelez, Kreiner, and Keiding (2010) for a graphical account of what population quantities are recoverable under outcome-dependent sampling.

6 See Griffith et al. (2020) for phenomena in data on COVID-19 patients that might be explained by collider bias.

for the result found in Lahtinen et al. (2019), who report that nonresponse bias leads to an underestimation of socioeconomic differences (X) in turnout (Y).[7]

This graphical structure also illustrates why explanations for nonresponse bias that are based only on correlations between observable variables (instead of graphs or structural equations) are insufficient. For example, Kennedy et al. (2018, p. 4) state that

> if survey response was correlated with presidential vote and some factor not accounted for in the weighting, then a deficient weighting protocol could be one explanation for the polling errors.

Note that the statement does not refer to causal order. In the collider structure, we have the correlation structure Kennedy et al. describe: survey response S correlates with candidate preference Y, and "some factor" X also correlates with S. However, this itself is not a problem – the collider structure and the conditioning on $S = 1$, however, is – and weighting or adjusting for X does not solve the problem. In fact, the bias introduced by weighting can be larger than the original bias.

Missingness at Random (MAR)

Sometimes survey research adjusts for differences between sample and population to make consistent inferences. Such an approach is valid if the outcome of interest is *missing at random* conditional on these adjustment variables, regardless of the specific method (post-stratification, propensity score weighting, etc.) used. We will now discuss a prototypical graph where such a strategy works before we return to the problem in more detail in Section 4.6. Figure 15 depicts a scenario where X are observed confounders—variables that affect both S and Y. In our running example, Kennedy et al. (2018) argue that a crucial element of X is education, which affected both survey participation and candidate preference.

In this case, Y is MAR because X d-separates the selection indicator from the outcome of interest. It is helpful to show

7 It is interesting to note that measurement error due to vote overreporting has the opposite effect since misreporters among nonvoters look like voters in terms of their SES Ansolabehere and Hersh, 2012, e.g.,

graphically why this is the case. To check for d-separation, we have to enumerate all paths between S and Y. There are two such paths: $S \leftarrow X \rightarrow Y$ and $S \leftarrow X \leftarrow U \rightarrow Y$. We see that in the first path, X is a confounder, while in the second path it is a mediator (between U and S). Accordingly, both paths are open: There is a correlation between sample selection and outcome running through X. After the 2016 presidential election, Kennedy et al. (2018) argued that education positively correlated with both sample selection and preferences for Hillary Clinton. Our graph suggests that this correlation might not only be because of causal effects of education itself, but also because there are deeper unobserved variables (e.g., early-life experiences) that also affect formal education and candidate preferences, but, crucially, do not directly affect survey participation.

In sum, as in the MNAR case, there is a correlation between survey participation and candidate preferences such that $P(Y) \neq P(Y|S = 1)$. However, in this graph, the dependence can be "broken" by conditioning on X. Doing so blocks the two paths connecting S and Y. Accordingly, we have $S \perp\!\!\!\perp Y|X$ and that $P(Y|X, S = 1) = P(Y|X)$. This means that the distribution of preferences as a function of education among sampled persons is the same as in the general population. Our analysis based on a substantively motivated graph and the use of d-separation shows why this works. In contrast, existing explanations rely on stating conditional independence assumptions coupled with informal explanations that are often imprecise, rather than deriving explanations from a deeper formal model. For example, Pfeffermann and Sverchkov (2009, p. 462) state that one needs to know "all the variables determining sample selection and response". In our graph, variables in U qualify for this criterion; however, we have shown that adjustment for them is not necessary.

If our interest is in the marginal distribution $P(Y)$, we can recover the population distribution of preferences if we know $P(X)$, the population distribution of education, perhaps from a census. We can then use the law of total probability and the assumption implied by the graph ($S \perp\!\!\!\perp Y|X$) to write:

$$\begin{aligned} P(Y) &= \sum_x P(Y|X = x)P(X = x) \\ &= \sum_x P(Y|X = x, S = 1)P(X = x). \end{aligned} \tag{63}$$

In the expression on the right-hand side, $P(Y|X = x, S = 1)$ can be estimated from our sample, and $P(X = x)$ is known from the census.

4.6 ADJUSTMENT FOR CONDITIONAL DISTRIBUTIONS

The preceding discussion has covered all relevant cases for recovering marginal distributions. We have also seen that one can recover conditional distributions when the conditioning variable (regressor) X is not only of substantive interest but also suffices to break the dependence between S and Y. The more typical case is one where the substantive interest is on $P(Y|X)$, but where it seems unlikely that X alone suffices for selection adjustment and additional variables Z are needed (see Figure 16 below). In this case, one cannot simply insert Z into the analysis (e.g., the regression model) because adjusting for it will usually change both the interpretation and the actual value of the regression coefficient of X. Curiously, we believe that this is a very common situation, yet it is never discussed in textbooks and has escaped the attention of the recent methodological literature (e.g., Elliott and Valliant, 2017). An exception is Pfeffermann and Sverchkov (2009), whose discussion is technical and not related to formal models of response behavior.

We saw earlier that if adjustment is possible, it involves adjusting for variables X that break the dependence between sample selection and outcome Y. If the interest is in a distribution already conditioned on X, but one thinks additional variables Z are needed, it is intuitive that identification is possible if X and Z together do the job to d-separate Y from S. In fact, a simple yet exhaustive graphical criterion for recoverability of $P(Y|X)$ is that (X, Z) d-separate S from Y, which implies $Y \perp\!\!\!\perp S|X, Z$.[8] Using this and the law of total probability, we can write (Bareinboim, Tian, and Pearl, 2014, p. 2413):

$$P(Y|X) = \sum_{Z=z} P(Y|X = x, Z = z, S = 1)P(Z = z|X = x). \quad (64)$$

In short, this expression asks us to compute the distribution (or mean) of Y given X and Z in the sample, and then to average

8 Although this technically is just a sufficient condition, and not necessary, it covers most scenarios where auxiliary variables are available (see Bareinboim, Tian, and Pearl, 2014).

over the X-specific population distribution of Z to obtain an estimate of the population distribution of Y conditional on X. This is a post-stratification estimator not for means, but for conditional distributions and means (see also Gelman, 2007). Under similar assumptions, one can also justify the use of standard inverse-probability weighting estimators, where the regression of interest is on X, and the weights are computed using X and Z. We refer the reader to Appendix 4.B for a deeper discussion of this topic.

4.7 MULTIPLE SELECTION NODES

So far, we have followed the literature on selection graphs and collapsed the data collection process into a single selection node S in order to understand the classical scenarios from the missingness literature. However, selecting respondents for surveys usually involves multiple stages. The TSE framework distinguishes between at least three of them: the *coverage stage* (1), in which members of the target population are selected into the frame population from which individuals are sampled at the *sampling stage* (2), and the *response stage* (3), in which sampled individuals further self-select according to their contactability, ability and willingness to participate in the survey. In each of these stages, inclusion may hinge on its own set of factors and can be related to the survey variable of interest in different ways. For instance, landline sampling frames are known for their lack of coverage of young and low-status persons (e.g., Blumberg and Luke, 2007); age and gender further affect the contactability of sampled subjects, presumably via their effects on employment status and mobility (e.g., Stoop, 2005); and social involvement as well as interactions between survey and respondent features, such as topic interest or trust in the sponsorship of a survey, are considered to be crucial for the cooperativeness among those successfully contacted (e.g., Groves, Singer, and Corning, 2000). In this section, we show that graphs are ideal for visualizing such multi-stage selection processes and that the common practice of using *design weights* only relies on very strong causal assumptions that are usually left implicit.

Multiple selection stages are easily represented in a graph by separate selection nodes. Consider the graphs in Figure 16, each of which includes two selection nodes, S_1 and S_2, for the

sampling and the response stage, respectively.[9] S_1 affects S_2 by design. Fundamentally, the obtained sample is conditional on both $S_1 = 1$ and $S_2 = 1$. The question, therefore, is whether the correlation between those two variables and the outcome of interest Y can be broken using an adjustment strategy. The left graph encodes the assumption that Z affects the selection of a unit in the sampling stage (S_1), but does not directly affect whether a sampled unit actually responds (S_2). Furthermore, Z affects the outcome. In this case, the analysis is very similar to the MAR case discussed previously: Z d-separates both selection variables from Y because it acts as a confounder on the paths between these variables and the outcome. Accordingly, the post-stratification formula $P(Y) = \sum_z P(Y|S_1 = 1, S_2 = 1, Z = z)P(Z = z)$ applies. In fact, for purposes of adjustment, we gain nothing by differentiating between S_1 and S_2.

Even though this graph may strike one as highly simplified and loaded with unrealistic assumptions concerning Z and (S_1, S_2), it is one of the weakest sets of assumptions that can be used to justify the common practice of "base" or "design weighting". For example, in telephone surveys, a sample of households is often drawn using random-digit dialing. On the first call, one person is selected so that individual inclusion probabilities at the sampling stage S_1 are inversely proportional to the number of eligible persons per household Z. This yields design weights $P(S_1 = 1|Z) = \frac{1}{Z}$. These are then used for estimating marginal or conditional distributions. Such an analysis disregards the response stage S_2 completely. In this graph, this adjustment strategy works because Z and S_1 d-separate S_2 from Y, so

$$\begin{aligned} P(Y) &= \sum_z P(Y|S_1 = 1, S_2 = 1, Z = z)P(Z = z) \\ &= \sum_z P(Y|S_1 = 1, Z = z)P(Z = z). \end{aligned} \tag{65}$$

It should be clear by now that such an approach does not yield consistent estimators if Z also affects whether units respond in the survey. This, unfortunately, seems quite likely (Gelman, 2007): Household size plausibly has a positive direct effect on contactibility and thereby S_2, consistent with the intermediate panel of Figure 16, while it has a negative effect

9 For ease of presentation, we assume that the frame population includes all members of the target population so that there is no selection through (under-)coverage.

(by design) on whether a unit is sampled in the first place. Formally, in this case, one would need to apply nonresponse weights based on $P(S_2|Z)$, as discussed previously.[10]

What are the consequences of unobserved confounders of the sampling and response stages, as in the right panel of Figure 16? Possession of a landline phone is an obvious variable that comes to mind: It clearly affects whether a unit may be sampled and may also have independent effects on individual response behavior S_2. In contrast to the left graph, response is now informative about the outcome Y, conditional on being sampled. This is because S_1 is a collider on the path $S_2 \leftarrow U \rightarrow S_1 \leftarrow Z \rightarrow Y$. However, Z still suffices as an adjustment variable. In fact, merely using design weights works. This is because Z d-separates S_2 from Y, conditional on S_1, so S_2 can be ignored. Even though S_1 is a collider that is conditioned upon, Z is a confounder on the same path, and so blocks it again.

As a final note, multiple selection nodes can also be used to depict item-specific missingness, which makes sense with sensitive survey items subject to substantial nonresponse. However, unit nonresponse has become much more prevalent than item nonresponse in recent years (see Yan and Curtin, 2010).[11]

4.8 DO ADJUSTMENT VARIABLES NEED TO CORRELATE WITH SURVEY SELECTION AND OUTCOME?

In the preceding discussion, we have emphasized that the assumptions for nonresponse adjustment implicitly or explicitly invoked are (conditional) independence assumptions—and only those. This is at odds with a sizable literature that investigates new adjustment variables from the survey process or administrative sources (Kreuter and Olson, 2011, 2013; Kreuter et al., 2010; Peytchev, Presser, and Zhang, 2018; Sakshaug and Antoni, 2019). In this literature, one finds different, or at least additional, requirements for valid adjustment variables. A representative statment reads, "to reduce bias effectively without increasing variance, a covariate that is used for nonre-

10 This is because $S_2 = 1$ implies $S_1 = 1$ logically, so the sampled data $P(Y|S_1 = 1, S_2 = 1$ can also be written as $P(Y|S_2 = 1)$. Therefore, weighting for S_2 is sufficient.

11 See Thoemmes and Mohan (2015) and Moreno-Betancur et al. (2018) for a graph theoretical account of item nonresponse.

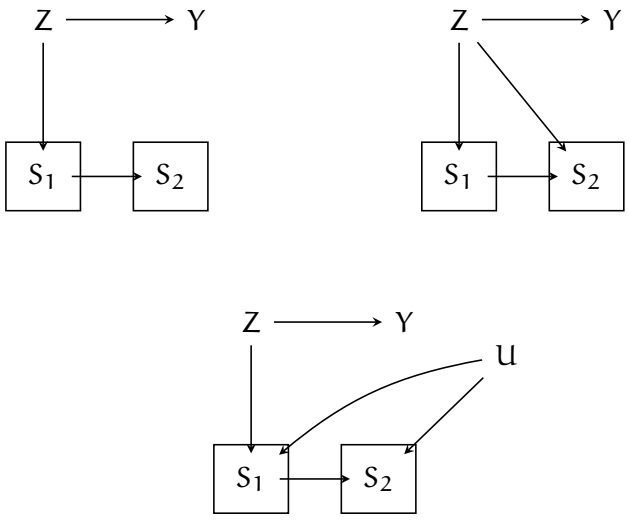

Figure 16: Recoverability with multi-stage selection: S_1 is selection through sampling, S_2 is selection at the response stage, X is a covariate, Y is a survey variable of interest, Z is an auxiliary variable, and U is an unobserved variable.

sponse weighting adjustment needs to be highly associated with both the response indicator and the survey outcome variable" (Kreuter et al., 2010, abstract).[12]

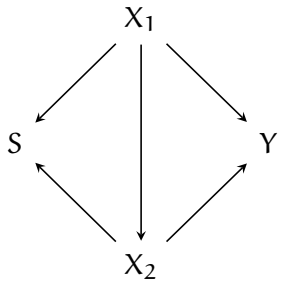

Figure 17: Graph where adjustment variables X_1 and X_2 may individually not correlate with Y at all due to offsetting paths, but the induced nonresponse bias may be large.

Using causal graphs, it is easy to see why this is not true. Consider the graph in Figure 17. X_1 could be age, while X_2 could be income (Kreuter and Olson, 2011, p. 316), and Y political preferences.

Here, the population correlation between X_1 and Y consists of the paths $X_1 \to Y$ and $X_1 \to X_2 \to Y$. These two paths may produce a very small or even zero correlation, e.g. when the direct effect of X_1 on Y is positive, but the the indirect effect

12 Similarly, Peytchev, Presser, and Zhang (2018) state in their abstract that such an association is a "*sine qua non* for effective adjustment" (emphasis in original).

is negative. On the other hand, the nonresponse bias via X_1 is due to the paths $S \leftarrow X_1 \rightarrow Y$ and $S \leftarrow X_1 \rightarrow X_2 \rightarrow Y$. Since here the effect $X_1 \rightarrow S$ is part of the paths, these could induce a completely different correlation, and introduce substantial nonresponse bias. (We show this in Appendix AX via a simple simulation.) Accordingly, we would want to adjust for X_1 (and X_2), even if its population correlation with Y was small.[13]

Kreuter and Olson (2011) analyze a very similar scenario using simulations, and in fact even illustrate it using a DAG. However, as far as we can see, they do not make the logical conclusion that inspecting correlations between adjustment variables and selection/outcome has no evidential value for whether the variables may be useful, and this may be one explanation for why it is still a wide-spread practice (Peytchev, Presser, and Zhang, 2018; Sakshaug and Antoni, 2019). Based on our analysis, we recommend researchers to regard correlations of adjustment variables with S and Y at most as descriptive information, and not as a basis for deciding which variables to use.

We note a final problem with these empirical studies, in that they confuse the in-sample correlation of X and Y with their population correlation. The preceding argument clearly relies on the population correlation, but the latter is not directly available to researchers, as Y is measured only for respondents. The in-sample correlation analyzed in the literature (Kreuter et al., 2010; Peytchev, Presser, and Zhang, 2018; Sakshaug and Antoni, 2019), on the other hand, may be badly biased for the population correlation. In the graph above, conditioning on S opens the path $X_1 \rightarrow S \leftarrow X_2 \rightarrow Y$. But this is not a constitutive part of the population correlation of X_1 and Y, and it will introduce bias. This is one more reason to disregard such sample-based estimates for diagnosis.

4.9 CONCLUSION

In this article, we have synthesized recent developments in the literature on causal graphs so that they are of great theoretical and practical value for survey researchers. By emphasizing the relationship between graphs and statistical dependencies, we have highlighted that graphs are tightly related to the statistical

13 This analysis immediately carries over to analyzing the correlation between X_1 and S.

machinery that survey researchers are already using. Insofar as their use has been informal (e.g., Groves, 2006), we have argued that they should be seen as formal mathematical models of the survey response process.

Specifically, causal graphs can be used to understand and communicate complex assumptions for nonresponse adjustment such as MAR and to derive general and straightforward estimators of marginal and associational parameters if certain assumptions are met. In Appendix 4.B, we show in detail that the independencies implied by a graph can be used to justify weighting estimators. Furthermore, multiple selection stages can easily be depicted using causal graphs, and we have shown the implicit assumptions behind using design weights. In this way, graphs allow for a unified view of random sampling and nonprobability samples, and design- and model-based survey inference (Elliott and Valliant, 2017; Kohler, Kreuter, and Stuart, 2019).

There are several interesting areas where future work can use causal graphs to visualize and better understand core assumptions in survey adjustment. One is the use of sample selection models when the assumptions for the adjustment strategies that we have discussed are not plausible. While early approaches (e.g., Heckman, 1979) relied on heavy parametric assumptions such as linearity in the structural models, more recent work relaxes these but emphasizes the importance of error independencies and exclusion restrictions (Das, Newey, and Vella, 2003). Both of these assumptions sets can be assessed using causal graphs.

We have not touched upon causal inference under sample selection here. Recent work on "transportability" (Bareinboim and Pearl, 2016) considers the related, but distinct task of generalizing causal effect estimates obtained from one study population to different populations. Transportability requires causal assumptions as well as non-experimental, observational data from the target population. For the latter, survey data will be indispensable.

In causal inference, another active area of research is the analysis of sensitivity to unobserved confounders. Mercer et al. (2017) call for similar developments with regard to sample selection bias. Recently, Smith and VanderWeele (2019) have de-

veloped a fairly general approach to this problem, guided by graphical assumptions.

The inference from self-reports to underlying attributes – measurement – is a further central task of any survey researcher. Graphs are excellent tools to depict complex measurement models and are widely used in conjunction with parametric assumptions in the structural equation modeling tradition (e.g., Veld and Saris, 2004). In the causal graph literature, weak assumptions on the sign and heterogeneity of the effects of latent constructs on measured quantities have been used to infer the existence of causal effects (VanderWeele and Hernán, 2012). We suspect that there is ample room for the development of empirical strategies for survey researchers using these approaches, especially with regard to the interaction of errors induced by sample selection and measurement error (Groves, 2006; Olson, 2019; Tourangeau, 2019).

Finally, we have not discussed the finite sample behavior – and especially the mean squared error – of various adjustment strategies. This is because we have been concerned with identification, which is the first logical step in nonresponse adjustment strategies. However, it is known that nonresponse adjustments may lead to estimates with high variance. This is usually investigated using Monte Carlo simulations, and research often relies on causal graphs to depict the structure behind the simulated data (Kreuter and Olson, 2011). There is some previous work in economics employing causal graphs to guide *efficient* covariate control for causal inference (White and Lu, 2011b) that might have implications for nonresponse adjustment as well.

APPENDIX

4.A COMPARISON TO THE ANALYSIS IN GROVES 2006

The three graphs in Figure 18 are slightly adapted from Groves, 2006, Figure 1, and superficially similar to the three prototypical selection graphs in Figure 15 that we have discussed. However, there are important differences. Groves (2006), in line with some of the survey literature on nonresponse (Bethlehem, 2002), does not work with a binary selection indicator S, but instead with an unobserved (latent) response propensity P, a number between 0 and 1 that characterizes every unit of the population

of interest. Nonresponse bias is introduced if P correlates with Y.

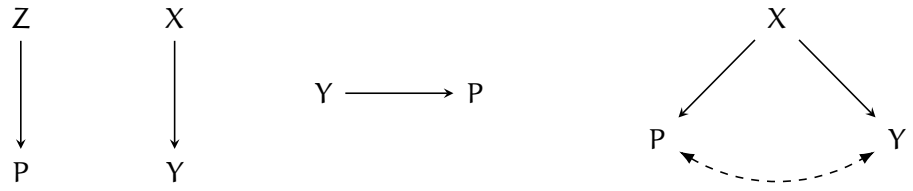

Figure 18: Three causal graphs, slightly adapted from Groves (2006, Figure 1). P stands for the "response propensity". The bidirected arrow in the right graph (presumably) indicates the association induced by X, not a separate unobserved confounder.

An advantage of using the binary S variable is that it consistent with the literature on partial identificaiton (Manski, 2009), and allows for the calculation of assumption-free bounds. Modelling the problem using P does not allow for such calculations, if only because the distribution of P is not known, not even for respondents.

Regarding the graphical models in Figure 18, the graph on the left is practically equivalent to the left graph in Figure 15. The center graphs, however, differ. It is only by introducing a second variable X also pointing into P or S that one is able to appreciate the collider bias phenomenon: Correlations may exist in-sample that are absent in the population. Again, it seems that the large literature on survey nonresponse has never explicitly discussed let alone explained this phenomenon. Finally, the graphs on the right also differ. In Figure 18, it seems (although it is not quite clear) that the additional bidirected path $P \leftrightarrow Y$ is meant to depict the consequence of having the common cause X, and not to visualize a separate unobserved confounder, as is common in the recent DAG literature (Pearl, 2009). This is not a trivial point. Using d-separation, we would conclude that successful nonresponse adjustment is not possible in this graph.

4.B INVERSE PROBABILITY WEIGHTING FOR M-ESTIMATION

To reiterate, we assume that (X, Z) block all paths from S to Y and that we have information on $P(Y, X, Z|S = 1)$ from the

sample and on $P(X, Z)$ from external data. Accordingly, we can assess the inclusion probabilities conditional on X and Z as

$$P(S = 1|X, Z) = \frac{P(X, Z|S = 1)P(S = 1)}{P(X, Z)}. \tag{66}$$

The left-hand side is easy to estimate using some regression approach with outcome S if X, Z are available for each observation in the sampling frame. If not, a natural solution is to calculate the expression on the right-hand side for all X, Z.

Our interest is to find an estimator for the parameter(s) β defined as

$$\min_{\beta} E[g(X, Y, \beta)], \tag{67}$$

where $g()$ is a specified function. For example, if X is a matrix of variables, y is a vector, and $X'X$ has full rank, then the population linear regression coefficients are $\beta = (X'X)^{-1}X'y$. Sample analogs of such minimization problems are called *M-estimators*. Maximum likelihood estimators are another special case of this very broad class of estimators.

This problem is different from the one considered in Wooldridge (2007). In that article, Z alone is sufficient for adjustment, and both X and Y can be seen as "outcomes". Here, we are interested in a situation where X is both needed for adjustment and as a conditioning variable of substantive interest, whereas Z is merely an auxiliary variable needed for adjustment.

Our approach will be to consider weighted expressions of the function $g()$ for some β (which we suppress for notational clarity) applied to the sampled observations, now indexed by individual i:

$$\frac{S_i g(X_i, Y_i)}{P(S_i = 1|X_i, Z_i)}. \tag{68}$$

To evaluate this expression, note that $P(S_i = 1|X_i, Z_i)$ is constant when X_i, Z_i are given. Also, $g(X_i, Y_i)$ is random only through Y_i when X_i, Z_i are given, so that under our causal assumptions $S_i \perp\!\!\!\perp g(X_i, Y_i)|X_i, Z_i$.[14] Using this, we have

14 The latter fact cannot be derived using d-separation.

$$
\begin{aligned}
E\left[\frac{S_i g(X_i, Y_i)}{P(S_i = 1|X_i, Z_i)}\right] &= E\left[E[\frac{S_i g(X_i, Y_i)}{P(S_i = 1|X_i, Z_i)}|X_i, Z_i]\right] \qquad (69)\\
&= E\left[\frac{1}{P(S_i = 1|X_i, Z_i)} E[S_i g(X_i, Y_i)|X_i, Z_i]\right] \\
&= E\left[\frac{1}{P(S_i = 1|X_i, Z_i)} E[S_i|X_i, Z_i] E[g(X_i, Y_i)|X_i, Z_i]\right] \\
&= E\left[\frac{1}{P(S_i = 1|X_i, Z_i)} P(S_i = 1|X_i, Z_i) E[g(X_i, Y_i)|X_i, Z_i]\right] \\
&= E[g(X_i, Y_i)].
\end{aligned}
$$

The first equality uses the law of iterated expectations. The second equality uses the fact that $P(S = 1)|X, Z$ is constant with respect to the inner expectation as well as linearity of expectations. The third equality uses the independence assumption $S \perp\!\!\!\perp g(X, Y)|X, Z$. The fourth equality follows because S is binary. The fifth equality again uses the law of iterated expectations. Under weak regularity conditions, minimizing the sample equivalent of equation 68 leads to an unbiased estimator of β (Wooldridge, 2007).

4.C HOW LOOKING AT CORRELATIONS CAN GO WRONG: A SIMULATION

Figure 19: Sampling distributions of correlation between X_1 and S (left) and of coefficient from a linear regression of Y on X_1 among respondents (right) across 1000 replications.

We use a simple and straightforward simulation setup similar to the ones in Kreuter and Olson (2011), and consistent with Figure 17:

$$X_1 \sim N(0, 1), \tag{70}$$

$$X_2 \sim N(X_1, 1), \tag{71}$$

$$S \sim Bern(p = logit^{-1}(2 + X_1 - X_2)), \tag{72}$$

$$Y \sim N(2 + X_1 - X_2, 2). \tag{73}$$

The interest in the population mean of Y. We simulate 1000 processes of the above model, each with 1000 observations in the population. The selection process means that Y is measured for 77.5% of the population, so the response probability is still quite high.

We evaluate the following diagnostics and estimators, in line with existing practice: First, we look at the correlation of X_1 and X_2 with S, based on the sampling frame (i.e., all 1000 observations). We then look at the regression of Y on X_1 or X_2 in the sample. Finally, we evaluate a weighting estimator for the mean of Y, based on inverse-probability weighting where we fit a logit model for S using both adjustment variables, or just the one that correlates with S or Y.

We start with adjustment variable X_2. On average, it correlates strongly with sample selection ($\rho \approx -0.48$), and is also strongly associated with Y in the sample ($\beta \approx -0.75$). This would lead us to correctly infer that it is an important adjustment variable.

However, Figure 19 plots the sampling distributions for the same statistics for adjustment variable X_1. Clearly, these are centered around zero. Most of the time, we would ignore X_1 as an adjustment variable, because its correlations with both S and Y would be exceedingly small.

However, when we evaluate the weighting estimators, the one based on both variables clearly outperforms the one based only on X_2. The latter has a relative bias of almost 10%, while the former is approximately unbiased. Even more impressively, using X_1 reduces mean squared error by 75%.

Accordingly, looking at correlations of adjustment variables with S or Y (the latter in sample) can lead analysts astray—in the extreme example here, one adjustment variable is completely uncorrelated with both S and Y (in sample), but still vastly improves nonresponse adjustment.

Part III

CONCLUSION

5

CONCLUSION

This dissertation has introduced the framework of causal graphs and structural causal models (Pearl, 2009) to political methodology and survey research, and has applied it to various problems in causal and statistical inference. In multiple instances, it has recommended researchers to avoid certain approaches or statistical tests but has also developed new tests and a sensitivity analysis for instrumental variables identification.

It has become clear that causal graphs are not only an intuitive tool to visualize assumptions, but can also be used to deduce testable and complicated counterfactual implication. A wide variety of problems and methods relevant for political science and many other disciplines—causal mediation, panel analysis, instrumental variables, and nonresponse adjustment—rely on such assumptions. A common theme in all chapters of this dissertation is that it is hard to understand conditional independence assumptions in isolation and that graphs are a natural framework to explain them.

In the following, I will conclude by discussing some contributions of this dissertation and their relationship to different literatures. I will also reflect on this dissertation's relationship to substantive theory and the so-called "credibility revolution".

5.1 BETWEEN NONPARAMETRIC AND PARAMETRIC ASSUMPTIONS

Although they were exclusively associated with linear causal models for decades, causal graphs are in some way inherently nonparametric: They visualize exclusion restrictions and the absence of unobserved confounders. These assumptions are unrelated to the functional form of relationships researchers may additionally assume. In most sciences, a stringent linearity assumption seems suspect. Paper 2 made some progress in this regard and developed a sensitivity analysis based on a semiparametric model. There are very powerful tools for causal

graphs based on linear causal models (Pearl, 2013), and there are recent developments on graphs based on monotonic (VanderWeele and Robins, 2010) or partially linear causal models (Rothenhäusler, Ernest, Bühlmann, et al., 2018). It may well be that these lead to interesting methods for the social sciences, especially in situations where there are multiple problems such as confounding, sample selection, and measurement error.

5.2 ROBUSTNESS TESTS

Paper 2 touched upon difficulties with "robustness" tests. Researchers compare two estimates, but it is often not clear which one, if any, we should believe if they differ. DAGs, on the other hand, automatically lead to all valid robustness tests based on d-separation.

More generally, Chen and Pearl (2015) remind us that for a valid robustness test one needs to have at least two sets of control variables (or, more generally, two different "methods") that are both plausible a priori and would identify an effect, and a chance that they give different answers if one or both of them is insufficient or false. This is a generalization of what the influential article by Hausman (1978) proposed, where one needs to maintain that at least one approach is consistent. In any case, the keywords here are "a priori" and "both plausible". In the cases discussed in paper 2, I find that the weaker set of assumptions—allowing for a direct effect of the instrument on M—were clearly more plausible. I think this is representative of most robustness tests that one sees: At the very least, the robustness test should have been the main specification. It would be interesting to look more deeply into what "effect stability" can realistically tell us. A viable, more transparent alternative seems to be sensitivity analysis, as developed in Paper 2. No observational causal inference is beyond doubt, and sensitivity analysis helps to put a number on how small or large this doubt may be.

5.3 EXTERNAL VALIDITY

The most recent development in causal graph theory was encompassing results on the possibility of "external validity" or "transportability" of causal effects (Bareinboim and Pearl, 2016;

Pearl and Bareinboim, 2014). Without making further parametric assumptions and simplifying matters a bit, this is possible if we can visualize the causes of differences (effect modifiers) between "source" and "target" population in a causal graph, and measure those causes in the source and target population. Importantly, social scientists will need to consider post-treatment effect modifiers. Although the tone in these publications is optimistic, a closer reading seems to cast doubt on the ability in the social sciences to produce highly generalizable knowledge, or at least on prior attempts to do so (e.g., Stuart et al., 2011). Social scientists will generally face situations where there are both pre- as well as post-treatment effect modifiers that need to be accounted for but are hard to measure. Consider information experiments: People with different priors will form different posteriors as a result of the information. If populations differ in their priors, this makes extrapolating effects hard; posteriors may be relevant post-treatment effect modifiers. Causal graphs will be instrumental in understanding this problem of external validity, and in developing new solutions to it.

5.4 SUBSTANTIVE THEORIES AND FORMAL MODELLING

Substantive theories played a somewhat small but central role in this dissertation. In paper 1, I emphasized the need to name substantive examples of unobserved confounders in a given research context, which I think would immediately improve most empirical research. I would generally think that most social science theories can be exhaustively depicted using a DAG, but this is also a problem. Saying "X influences Y either through M or through W" is a very superficial theory. I would hope that DAGs, which are formal models that can be used to point to and explain counterintuitive phenomena like collider bias, convey to skeptics the more general appeal of formal modeling, and push the social sciences into this direction also on the theory front.

5.5 THE CREDIBILITY REVOLUTION

The approach followed in this dissertation has an interesting relation to the "credibility revolution" (Angrist and Pischke, 2010), or what Samii calls "causal empiricism" (Samii, 2016).

According to Samii, the focus here is on nonparametric identification as well as the use of experiments and "natural" experiments to establish "specific causal facts for well-defined subpopulations" (Samii, 2016, p. 941). I would add that an essential ingredient for "natural" experiments is detailed knowledge of the subject at hand, e.g., institutional rules that are exploited in regression-discontinuity designs.

While this dissertation was all about nonparametric identification and the closely related task to define causal effects, it is not clear how this is related to higher standards for identification (as in: is there really no direct effect of the instrument?), or the more detailed historical or institutional knowledge quantitative researchers in the social sciences use nowadays. But then again, it is also not clear (to me) how this is related to the potential outcomes approach advocated by methodologists associated with the credibility revolution.

For example, Paper 2 showed that many researchers continue to have a somewhat vague understanding of the necessary assumptions for instrumental variables estimation, at least outside of the ideal, randomized, three variables case. While I do not doubt that social scientists nowadays are, for example, much more sensitive to finding regression discontinuities or potential instruments ("this sounds like it's an instrument!"), I am doubtful that a focus on potential outcomes and nonparametric identification has had a positive causal effect on this.

An interesting piece of evidence was (accidentally) provided by Imbens in his discussion of the potential outcomes vis-à-vis the "DAG approach" (Imbens, 2019). The final sentences of that paper are (Imbens, 2019, p. 57):

> It is studies such as LaLonde (1986), Card (1990), Angrist (1990), Angrist and Krueger (1991), and Ashenfelter and Krueger (1994) that spurred the credibility revolution and the adoption of the PO framework, not the theoretical advances. There have not been similar applications of the DAG framework, and more papers discussing toy models will not be sufficient.

But none of these papers contain potential outcomes notation. In fact, LaLonde (1986), Angrist (1990), Angrist and Krueger (1991) and Ashenfelter and Krueger (1994) all contain treatments of good

old-fashioned linear structural models of varying complexity, while Card (1990) contains almost no mathematical notation at all. Finally, I note that all DAGs in this dissertation could be described as "toy model", i.e., models with just around a handful of variables.

This is representative of the strange pirouettes that the discussion around "the potential outcomes versus the DAG approach" has performed. The distinction has no factual scientific basis to begin with. I hope this dissertation has made this clear, and that it contributes to opening up the social sciences to engage in more systematic and transparent discussions about empirical research.

BIBLIOGRAPHY

Abramson, Scott F (2017). "The economic origins of the territorial state." In: *International Organization* 71.1, pp. 97–130.

Acharya, Avidit, Matthew Blackwell, and Maya Sen (2016). "Explaining causal findings without bias: Detecting and assessing direct effects." In: *American Political Science Review* 110.3, pp. 512–529.

Ahmed, Faisal Z (2012). "The perils of unearned foreign income: Aid, remittances, and government survival." In: *American Political Science Review* 106.1, pp. 146–165.

Albert, Jeffrey M and Wei Wang (2015). "Sensitivity analyses for parametric causal mediation effect estimation." In: *Biostatistics* 16.2, pp. 339–351.

Angrist, Joshua D (1990). "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." In: *American Economic Review* 80.3, pp. 313–336. URL: https://ideas.repec.org/a/aea/aecrev/v80y1990i3p313-36.html.

Angrist, Joshua D and Guido W Imbens (1995). "Two-stage least squares estimation of average causal effects in models with variable treatment intensity." In: *Journal of the American statistical Association* 90.430, pp. 431–442.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin (1996). "Identification of causal effects using instrumental variables." In: *Journal of the American statistical Association* 91.434, pp. 444–455.

Angrist, Joshua D and Alan B Krueger (1991). "Does compulsory school attendance affect schooling and earnings?" In: *The Quarterly Journal of Economics* 106.4, pp. 979–1014.

Angrist, Joshua D and Jörn-Steffen Pischke (2010). "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." In: *Journal of economic perspectives* 24.2, pp. 3–30.

Angrist, Joshua David and Jörn-Steffen Pischke (2009). *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press.

Angrist, Joshua and Ivan Fernandez-Val (2013). "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." In: *Advances in Economics and Econometrics: Theory and Ap-*

plications, Tenth World Congress, Volume III: Econometrics. Econometric Society Monographs.

Ansolabehere, Stephen and Eitan Hersh (2012). "Validation: What big data reveal about survey misreporting and the real electorate." In: *Political Analysis* 20.4, pp. 437–459.

Aronow, Peter M and Allison Carnegie (2013). "Beyond LATE: Estimation of the average treatment effect with an instrumental variable." In: *Political Analysis* 21.4, pp. 492–506.

Aronow, Peter M, Donald P Green, Donald KK Lee, et al. (2014). "Sharp bounds on the variance in randomized experiments." In: *The Annals of Statistics* 42.3, pp. 850–871.

Aronow, Peter M and Benjamin T Miller (2019). *Foundations of agnostic statistics*. Cambridge University Press.

Aronow, Peter M, Cyrus Samii, et al. (2017). "Estimating average causal effects under general interference, with application to a social network experiment." In: *The Annals of Applied Statistics* 11.4, pp. 1912–1947.

Arzheimer, Kai (2009). "Contextual factors and the extreme right vote in Western Europe, 1980–2002." In: *American Journal of Political Science* 53.2, pp. 259–275.

Ashenfelter, Orley and Alan Krueger (1994). "Estimates of the Economic Return to Schooling from a New Sample of Twins." In: *The American Economic Review* 84.5, pp. 1157–1173.

Bareinboim, Elias and Judea Pearl (2016). "Causal inference and the data-fusion problem." In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7345–7352.

Bareinboim, Elias, Jin Tian, and Judea Pearl (2014). "Recovering from selection bias in causal and statistical inference." In: *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Bazzi, Samuel and Michael A Clemens (2013). "Blunt instruments: avoiding common pitfalls in identifying the causes of economic growth." In: *American Economic Journal: Macroeconomics* 5.2, pp. 152–186.

Bellemare, Marc F, Takaaki Masaki, and Thomas B Pepinsky (2017). "Lagged explanatory variables and the estimation of causal effect." In: *The Journal of Politics* 79.3, pp. 949–963.

Berinsky, Adam J and Sara Chatfield (2015). "An empirical justification for the use of draft lottery numbers as a random treatment in political science research." In: *Political Analysis* 23.3, pp. 449–454.

Bethlehem, Jelke (2002). "Weighting Nonresponse Adjustments Based on Auxiliary Information." en. In: *Survey Nonresponse*. Ed. by

Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J.A. Little. New York: Wiley, 275–88.

Betz, Timm, Scott J Cook, and Florian M Hollenbach (2018). "On the use and abuse of spatial instruments." In: *Political Analysis*, pp. 1–6.

Blackwell, Matthew and Adam N Glynn (2018). "How to make causal inferences with time-series cross-sectional data under selection on observables." In: *American Political Science Review* 112.4, pp. 1067–1082.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys (2019). "Declaring and diagnosing research designs." In: *American Political Science Review* 113.3, pp. 838–859.

Blumberg, Stephen J and Julian V Luke (2007). "Coverage bias in traditional telephone surveys of low-income and young adults." In: *Public Opinion Quarterly* 71.5, pp. 734–749.

Boix, Carles (2011). "Democracy, development, and the international system." In: *American Political Science Review* 105.4, pp. 809–828.

Brand, Jennie E and Yu Xie (2010). "Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education." In: *American sociological review* 75.2, pp. 273–302.

Card, David (1990). "The impact of the Mariel boatlift on the Miami labor market." In: *ILR Review* 43.2, pp. 245–257.

— (1999). "The causal effect of education on earnings." In: *Handbook of labor economics*. Vol. 3. Elsevier, pp. 1801–1863.

Carnegie, Allison and Nikolay Marinov (2017). "Foreign Aid, Human Rights, and Democracy Promotion: Evidence from a Natural Experiment." In: *American Journal of Political Science*.

Caro, Daniel H (2015). "Causal mediation in educational research: An illustration using international assessment data." In: *Journal of Research on Educational Effectiveness* 8.4, pp. 577–597.

Chen, Bryant and Judea Pearl (2013). "Regression and causation: a critical examination of six econometrics textbooks." In: *Real-World Economics Review, Issue* 65, pp. 2–20.

— (2015). *Exogeneity and robustness*. Tech. rep. Tech. Rep.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey (2013). "Average and quantile effects in nonseparable panel models." In: *Econometrica* 81.2, pp. 535–580.

Cinelli, Carlos and Chad Hazlett (2020). "Making sense of sensitivity: Extending omitted variable bias." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1, pp. 39–67.

Cinelli, Carlos and Judea Pearl (2018). "RE: a practical example demonstrating the utility of single-world intervention graphs." In: *Epidemiology* 29.6, e50–e51.

Claassen, Christopher (2020). "Does public support help democracy survive?" In: *American Journal of Political Science* 64.1, pp. 118–134.

Conley, Timothy G, Christian B Hansen, and Peter E Rossi (2012). "Plausibly exogenous." In: *Review of Economics and Statistics* 94.1, pp. 260–272.

Coppock, Alexander et al. (2019). "Avoiding post-treatment bias in audit experiments." In: *Journal of Experimental Political Science* 6.1, pp. 1–4.

Cornesse, Carina, Annelies G Blom, David Dutwin, Jon A Krosnick, Edith D De Leeuw, Stéphane Legleye, Josh Pasek, Darren Pennay, Benjamin Phillips, Joseph W Sakshaug, et al. (2020). "A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research." In: *Journal of Survey Statistics and Methodology*.

Daniel, Rhian M, Michael G Kenward, Simon N Cousens, and Bianca L De Stavola (2012). "Using causal diagrams to guide analysis in missing data problems." In: *Statistical Methods in Medical Research* 21.3, pp. 243–256.

Das, Mitali, Whitney K Newey, and Francis Vella (2003). "Nonparametric estimation of sample selection models." In: *The Review of Economic Studies* 70.1, pp. 33–58.

Dawid, A Philip (1979). "Conditional independence in statistical theory." In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–31.

De Mesquita, Ethan Bueno and Scott A Tyson (2020). "The commensurability problem: Conceptual difficulties in estimating the effect of behavior on behavior." In: *American Political Science Review* 114.2, pp. 375–391.

Deuchert, Eva and Martin Huber (2017). "A cautionary tale about control variables in IV estimation." In: *Oxford Bulletin of Economics and Statistics* 79.3, pp. 411–425.

Didelez, Vanessa, Svend Kreiner, and Niels Keiding (2010). "Graphical models for inference under outcome-dependent sampling." In: *Statistical Science* 25.3, pp. 368–387.

Elliott, Michael R and Richard Valliant (2017). "Inference for nonprobability samples." In: *Statistical Science* 32.2, pp. 249–264.

Ellis, Bret Easton (1991). *American Psycho: a novel.* Vintage Books.

Elwert, Felix and Christopher Winship (2014a). "Endogenous selection bias: The problem of conditioning on a collider variable." In: *Annual Review of Sociology* 40, pp. 31–53.

— (2014b). "Endogenous selection bias: The problem of conditioning on a collider variable." In: *Annual Review of Sociology* 40, pp. 31–53.

Esterling, Kevin M, Michael A Neblo, and David MJ Lazer (2011). "Estimating treatment effects in the presence of noncompliance and nonresponse: The generalized endogenous treatment model." In: *Political Analysis*, pp. 205–226.

Frölich, Markus and Martin Huber (2017). "Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. ISSN: 1467-9868. DOI: 10.1111/rssb.12232. URL: http://dx.doi.org/10.1111/rssb.12232.

Geiger, Dan, Thomas Verma, and Judea Pearl (1990). "Identifying independence in Bayesian networks." In: *Networks* 20.5, pp. 507–534.

Gelman, Andrew (2007). "Struggles with survey weighting and regression modeling." In: *Statistical Science* 22.2, pp. 153–164.

Gelman, Andrew, Sharad Goel, Douglas Rivers, David Rothschild, et al. (2016). "The mythical swing voter." In: *Quarterly Journal of Political Science* 11.1, pp. 103–130.

Glynn, Adam N (2012). "The product and difference fallacies for indirect effects." In: *American Journal of Political Science* 56.1, pp. 257–269.

Glynn, Adam N. and Konstantin Kashin (2017). "Front-Door Difference-in-Differences Estimators." In: *American Journal of Political Science*.

Greenland, Sander, Judea Pearl, and James M Robins (1999). "Causal diagrams for epidemiologic research." In: *Epidemiology* 10.1, pp. 37–48.

Griffith, Gareth et al. (2020). "Collider bias undermines our understanding of COVID-19 disease risk and severity." In: *medRxiv*. DOI: 10.1101/2020.05.04.20090506. URL: https://www.medrxiv.org/content/early/2020/05/08/2020.05.04.20090506.

Groves, Robert M (2006). "Nonresponse rates and nonresponse bias in household surveys." In: *Public Opinion Quarterly* 70.5, pp. 646–675.

Groves, Robert M, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau (2009). *Survey methodology*. 2nd ed. John Wiley & Sons.

Groves, Robert M and Lars Lyberg (2010). "Total survey error: Past, present, and future." In: *Public Opinion Quarterly* 74.5, pp. 849–879.

Groves, Robert M and Emilia Peytcheva (2008). "The impact of nonresponse rates on nonresponse bias: a meta-analysis." In: *Public Opinion Quarterly* 72.2, pp. 167–189.

Groves, Robert M, Stanley Presser, and Sarah Dipko (2004). "The role of topic interest in survey participation decisions." In: *Public Opinion Quarterly* 68.1, pp. 2–31.

Groves, Robert M, Eleanor Singer, and Amy Corning (2000). "Leverage-saliency theory of survey participation: description and an illustration." In: *The Public Opinion Quarterly* 64.3, pp. 299–308.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw (2001). "Identification and estimation of treatment effects with a regression-discontinuity design." In: *Econometrica* 69.1, pp. 201–209.

Hainmueller, Jens, Daniel J Hopkins, and Teppei Yamamoto (2014). "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." In: *Political analysis* 22.1, pp. 1–30.

Hausman, Jerry A (1978). "Specification tests in econometrics." In: pp. 1251–1271.

Heckman, James J (1979). "Sample selection bias as a specification error." In: *Econometrica*, pp. 153–161.

Heckman, James J and Rodrigo Pinto (2018). "Unordered monotonicity." In: *Econometrica* 86.1, pp. 1–35.

Heckman, James and Rodrigo Pinto (2015). "CAUSAL ANALYSIS AFTER HAAVELMO." In: *Econometric Theory* 31.1, p. 115.

Holland, Paul W (1986). "Statistics and causal inference." In: *Journal of the American statistical Association* 81.396, pp. 945–960.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppeiarlarl Yamamoto (2011). "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies." In: *American Political Science Review* 105.4, pp. 765–789.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto (2010). "Identification, inference and sensitivity analysis for causal mediation effects." In: *Statistical Science* 25.1, pp. 51–71.

Imai, Kosuke and In Song Kim (2019). "When should we use unit fixed effects regression models for causal inference with longitudinal data?" In: *American Journal of Political Science* 63.2, pp. 467–490.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto (2013). "Experimental designs for identifying causal mechanisms." In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176.1, pp. 5–51.

Imai, Kosuke and Teppei Yamamoto (2013). "Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments." In: *Political Analysis* 21.2, pp. 141–171.

Imbens, Guido W (2010). "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." In: *Journal of Economic literature* 48.2, pp. 399–423.

— (2014). "Instrumental Variables: An Econometrician's Perspective." In: *Statistical Science* 29.3, pp. 323–358.

Imbens, Guido W and Whitney K Newey (2009). "Identification and estimation of triangular simultaneous equations models without additivity." In: *Econometrica* 77.5, pp. 1481–1512.

Imbens, Guido W and Donald B Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, Guido (2019). *Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics*. Tech. rep. National Bureau of Economic Research.

Jackman, Robert W and Karin Volpert (1996). "Conditions favouring parties of the extreme right in Western Europe." In: *British Journal of Political Science* 26.4, pp. 501–521.

Kafka, Franz (1954). "Tagebücher, 1910-1923." In: *Gesammelte Werke*. S. Fischer.

Keele, Luke (2015a). "Causal Mediation Analysis Warning! Assumptions Ahead." In: *American Journal of Evaluation* 36.4, pp. 500–513.

— (2015b). "The statistics of causal inference: A view from political methodology." In: *Political Analysis* 23.3, pp. 313–335.

Keele, Luke, Randolph T Stevenson, and Felix Elwert (2020). "The causal interpretation of estimated associations in regression models." In: *Political Science Research and Methods* 8.1, pp. 1–13.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers, et al. (2018). "An evaluation of the 2016 election polls in the United States." In: *Public Opinion Quarterly* 82.1, pp. 1–33.

Kern, Holger Lutz and Jens Hainmueller (2009). "Opium for the masses: How foreign media can stabilize authoritarian regimes." In: *Political Analysis* 17.4, pp. 377–399.

King, Gary (1998). *Unifying political methodology: The likelihood theory of statistical inference*. University of Michigan Press.

King, Gary, Ori Rosen, Martin Tanner, and Alexander F Wagner (2008). "Ordinary economic voting behavior in the extraordinary election of Adolf Hitler." In: *The Journal of Economic History* 68.4, pp. 951–996.

Knox, Dean, Will Lowe, and Jonathan Mummolo (2020). "Administrative Records Mask Racially Biased Policing." In: *American Political Science Review*, 1–19.

Kocher, Matthew Adam, Thomas B Pepinsky, and Stathis N Kalyvas (2011). "Aerial bombing and counterinsurgency in the Vietnam War." In: *American Journal of Political Science* 55.2, pp. 201–218.

Kohler, Ulrich, Frauke Kreuter, and Elizabeth A Stuart (2019). "Non-probability sampling and causal analysis." In: *Annual Review of Statistics and its Application* 6, pp. 149–172.

Kreuter, Frauke and Kristen Olson (2011). "Multiple auxiliary variables in nonresponse adjustment." In: *Sociological Methods & Research* 40.2, pp. 311–332.

— (2013). "Paradata for nonresponse error investigation." In: *Improving surveys with paradata: Analytic uses of process information* 2, pp. 13–42.

Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trena M Ezzati-Rice, Carolina Casas-Cordero, Michael Lemay, Andy Peytchev, Robert M Groves, and Trivellore E Raghunathan (2010). "Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys." In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173.2, pp. 389–407.

LaLonde, Robert J (1986). "Evaluating the econometric evaluations of training programs with experimental data." In: *The American economic review*, pp. 604–620.

Lahtinen, Hannu, Pekka Martikainen, Mikko Mattila, Hanna Wass, and Lauri Rapeli (2019). "Do Surveys Overestimate or Underestimate Socioeconomic Differences in Voter Turnout? Evidence from Administrative Registers." In: *Public Opinion Quarterly*.

Lee, David S and Thomas Lemieux (2010). "Regression discontinuity designs in economics." In: *Journal of economic literature* 48.2, pp. 281–355.

Lee, Youjin and Elizabeth L Ogburn (2020). "Network Dependence Can Lead to Spurious Associations and Invalid Inference." In: *Journal of the American Statistical Association* just-accepted, pp. 1–31.

Little, Roderick JA and Donald B Rubin (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Loeys, Tom, Beatrijs Moerkerke, Olivia De Smet, Ann Buysse, Johan Steen, and Stijn Vansteelandt (2013). "Flexible mediation analysis in the presence of nonlinear relations: beyond the mediation formula." In: *Multivariate Behavioral Research* 48.6, pp. 871–894.

Lohr, Sharon L (2019). *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Lu, Xun and Halbert White (2014). "Robustness checks and robustness tests in applied economics." In: *Journal of econometrics* 178, pp. 194–206.

Maathuis, Marloes H and Diego Colombo (2015). "A generalized back-door criterion." In: *The Annals of Statistics* 43.3, pp. 1060–1088.

Maclaren, Oliver J and Ruanui Nicholson (2019). "What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems." In: *arXiv preprint arXiv:1904.02826*.

Manski, Charles F (2009). *Identification for prediction and decision*. Harvard University Press.

Marshall, John (2016). "Coarsening Bias: How Coarse Treatment Measurement Upwardly Biases Instrumental Variable Estimates." In: *Political Analysis* 24.2, pp. 157–171.

Mercer, Andrew W, Frauke Kreuter, Scott Keeter, and Elizabeth A Stuart (2017). "Theory and practice in nonprobability surveys: parallels between causal inference and survey inference." In: *Public Opinion Quarterly* 81.S1, pp. 250–271.

Mohan, Karthika and Judea Pearl (2019). "Graphical models for processing missing data." In: *Journal of American Statistical Association*.

Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres (2018). "How conditioning on posttreatment variables can ruin your experiment and what to do about it." In: *American Journal of Political Science* 62.3, pp. 760–775.

Moreno-Betancur, Margarita, Katherine J Lee, Finbarr P Leacy, Ian R White, Julie A Simpson, and John B Carlin (2018). "Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies." In: *American journal of epidemiology* 187.12, pp. 2705–2715.

Morgan, Stephen L and Christopher Winship (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Naimi, Ashley I, Jay S Kaufman, and Richard F MacLehose (2014). "Mediation misgivings: Ambiguous clinical and public health in-

terpretations of natural direct and indirect effects." In: *International Journal of Epidemiology* 43.5, pp. 1656–1661.

Nutz, Marcel and Ruodu Wang (2020). "The Directional Optimal Transport." In: arXiv: 2002.08717 [math.OC].

Ogburn, Elizabeth L, Tyler J VanderWeele, et al. (2014). "Causal diagrams for interference." In: *Statistical science* 29.4, pp. 559–578.

Olson, Kristen (2019). "Comments On "How Errors Cumulate: Two Examples" by Roger Tourangeau." In: *Journal of Survey Statistics and Methodology*.

Pearl, Judea (1993). "[Bayesian analysis in expert systems]: comment: graphical models, causality and intervention." In: *Statistical Science* 8.3, pp. 266–269.

— (2001). "Direct and indirect effects." In: *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 411–420.

— (2009). *Causality*. Cambridge university press.

— (2013). "Linear models: A useful "microscope" for causal analysis." In: *Journal of Causal Inference* 1.1, pp. 155–170.

— (2014). "Interpretation and identification of causal mediation." In: *Psychological Methods* 19.4, p. 459.

— (2016). *Recollections from the WCE conference at Stanford*. URL: http://causality.cs.ucla.edu/blog/index.php/2016/06/20/recollections-from-the-wce-conference-at-stanford/.

Pearl, Judea and Elias Bareinboim (2014). "External validity: From do-calculus to transportability across populations." In: *Statistical Science*, pp. 579–595.

Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt (2017). "Poorly Measured Confounders are More Useful on the Left Than on the Right." In: *NBER Working Paper* 23232.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference*. The MIT Press.

Peytchev, Andy (2013). "Consequences of survey nonresponse." In: *The ANNALS of the American Academy of Political and Social Science* 645.1, pp. 88–111.

Peytchev, Andy, Lisa R Carley-Baxter, and Michele C Black (2011). "Multiple sources of nonobservation error in telephone surveys: coverage and nonresponse." In: *Sociological Methods & Research* 40.1, pp. 138–168.

Peytchev, Andy, Stanley Presser, and Mengmeng Zhang (2018). "Improving traditional nonresponse bias adjustments: Combining statistical properties with social theory." In: *Journal of Survey Statistics and Methodology* 6.4, pp. 491–515.

Pfeffermann, Danny and Michail Sverchkov (2009). "Inference under informative sampling." In: *Handbook of statistics*. Vol. 29. Elsevier, pp. 455–487.

Pierskalla, Jan H and Florian M Hollenbach (2013). "Technology and collective action: The effect of cell phone coverage on political violence in Africa." In: *American Political Science Review* 107.02, pp. 207–224.

Powell, James L (1994). "Estimation of semiparametric models." In: *Handbook of econometrics* 4, pp. 2443–2521.

Richardson, Thomas S and James M Robins (2013). "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality." In:

Rosenbaum, Paul R (1984). "The consquences of adjustment for a concomitant variable that has been affected by the treatment." In: *Journal of the Royal Statistical Society. Series A (General)*, pp. 656–666.

Rothenhäusler, Dominik, Jan Ernest, Peter Bühlmann, et al. (2018). "Causal inference in partially linear structural equation models." In: *The Annals of Statistics* 46.6A, pp. 2904–2938.

Rubin, Donald B (1976). "Inference and missing data." In: *Biometrika* 63.3, pp. 581–592.

— (2009). "Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?" In: *Statistics in Medicine* 28.9, pp. 1420–1423.

Sakshaug, Joseph W and Manfred Antoni (2019). "Evaluating the utility of indirectly linked federal administrative records for nonresponse bias adjustment." In: *Journal of Survey Statistics and Methodology* 7.2, pp. 227–249.

Samii, Cyrus (2016). "Causal empiricism in quantitative research." In: *The Journal of Politics* 78.3, pp. 941–955.

Samii, Cyrus, Laura Paler, and Sarah Zukerman Daly (2016). "Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia." In: *Political Analysis* 24.4, pp. 434–456.

Schafer, Joseph L and John W Graham (2002). "Missing data: our view of the state of the art." In: *Psychological methods* 7.2, p. 147.

Schneider, Carsten Q (2018). "Realists and Idealists in QCA." In: *Political Analysis* 26.2, pp. 246–254.

Schultz, Kenneth A and Justin S Mankin (2019). "Is temperature exogenous? The impact of civil conflict on the instrumental climate record in Sub-Saharan Africa." In: *American Journal of Political Science* 63.4, pp. 723–739.

Shalizi, Cosma (2019). *Advanced data analysis from an elementary point of view*. URL: http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf.

Shirani-Mehr, Houshmand, David Rothschild, Sharad Goel, and Andrew Gelman (2018). "Disentangling bias and variance in election polls." In: *Journal of the American Statistical Association* 113.522, pp. 607–614.

Shpitser, Ilya, Tyler VanderWeele, and James M Robins (2010). "On the validity of covariate adjustment for estimating causal effects." In: *26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*, pp. 527–536.

Simon, Herbert A (1954). "Bandwagon and underdog effects and the possibility of election predictions." In: *Public Opinion Quarterly* 18.3, pp. 245–253.

Smith, Louisa H and Tyler J VanderWeele (2019). "Bounding bias due to selection." In: *Epidemiology* 30.4, pp. 509–516.

Sovey, Allison J and Donald P Green (2011). "Instrumental variables estimation in political science: A readers' guide." In: *American Journal of Political Science* 55.1, pp. 188–200.

Spenkuch, Jörg L and Philipp Tillmann (2017). "Elite Influence? Religion and the Electoral Success of the Nazis." In: *American Journal of Political Science*.

Spiegler, Ran (2016). "Bayesian networks and boundedly rational expectations." In: *The Quarterly Journal of Economics* 131.3, pp. 1243–1290.

Spiegler, Ran and Kfir Eliaz (Forthcoming). "A Model of Competing Narratives." In: *American Economic Review*.

Stanig, Piero (2015). "Regulation of speech and media coverage of corruption: An empirical analysis of the Mexican Press." In: *American Journal of Political Science* 59.1, pp. 175–193.

Stoop, Ineke AL (2005). *The hunt for the last respondent: Nonresponse in sample surveys*. Vol. 200508. Sociaal en Cultureel Planbu.

Stuart, Elizabeth A, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf (2011). "The use of propensity scores to assess the generalizability of results from randomized trials." In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174.2, pp. 369–386.

Textor, Johannes, Juliane Hardt, and Sven Knüppel (2011). "DAGitty: a graphical tool for analyzing causal diagrams." In: *Epidemiology* 22.5, p. 745.

Textor, Johannes, Benito van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison (2016). "Robust causal infer-

ence using directed acyclic graphs: the R package 'dagitty'." In: *International journal of epidemiology* 45.6, pp. 1887–1894.

Thoemmes, Felix and Karthika Mohan (2015). "Graphical representation of missing data problems." In: *Structural Equation Modeling: A Multidisciplinary Journal* 22.4, pp. 631–642.

Tourangeau, Roger (2019). "How Errors Cumulate: Two Examples." In: *Journal of Survey Statistics and Methodology*.

Trounstine, Jessica (2016). "Segregation and inequality in public goods." In: *American Journal of Political Science* 60.3, pp. 709–725.

Valliant, Richard, Alan H. Dorfman, and Richard M. Royall (2000). *Finite population sampling and inference. A prediction approach*. John Wiley & Sons.

Van Kippersluis, Hans and Cornelius A Rietveld (2018). "Pleiotropy-robust Mendelian randomization." In: *International Journal of Epidemiology* 47.4, pp. 1279–1288.

VanderWeele, Tyler J and Miguel A Hernán (2012). "Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs." In: *American Journal of Epidemiology* 175.12, pp. 1303–1310.

VanderWeele, Tyler J and James M Robins (2010). "Signed directed acyclic graphs for causal inference." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.1, pp. 111–127.

Veld, William van der and Willem E. Saris (2004). "Separation of Error, Method Effects, Instability, and Attitude Strength." In: *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change*. Ed. by Willem E. Saris and Paul M. Sniderman. Princeton, NJ: Princeton University Press, pp. 37–59.

Wasserman, Larry (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

White, Ariel (2019). "Misdemeanor Disenfranchisement? The demobilizing effects of brief jail spells on potential voters." In: *American Political Science Review* 113.2, pp. 311–324.

White, Halbert and Karim Chalak (2010). "Testing a conditional form of exogeneity." In: *Economics Letters* 109.2, pp. 88–90.

White, Halbert and Xun Lu (2011a). "Causal diagrams for treatment effect estimation with application to efficient covariate selection." In: *Review of Economics and Statistics* 93.4, pp. 1453–1459.

— (2011b). "Causal diagrams for treatment effect estimation with application to efficient covariate selection." In: *Review of Economics and Statistics* 93.4, pp. 1453–1459.

Woodberry, Robert D (2012). "The missionary roots of liberal democracy." In: *American Political Science Review* 106.2, pp. 244–274.

Wooldridge, Jeffrey M (2007). "Inverse probability weighted estimation for general missing data problems." In: *Journal of econometrics* 141.2, pp. 1281–1301.

— (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wright, Sewall (1920). "The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs." In: *Proceedings of the National Academy of Sciences of the United States of America* 6.6, p. 320.

Wucherpfennig, Julian, Philipp Hunziker, and Lars-Erik Cederman (2016). "Who inherits the state? Colonial rule and postcolonial conflict." In: *American Journal of Political Science* 60.4, pp. 882–898.

Yan, Ting and Richard Curtin (2010). "The relation between unit nonresponse and item nonresponse: A response continuum perspective." In: *International Journal of Public Opinion Research* 22.4, pp. 535–551.

Zhang, Li-Chun (2000). "Post-stratification and calibration—a synthesis." In: *The American Statistician* 54.3, pp. 178–184.

Part IV

AUTHOR'S CONTRIBUTION

AUTHOR'S CONTRIBUTION

Chapter 2: Causal Graphs in Political Methodology (Chapter 2)

The article is single-authored.

Chapter 3: Post-Instrument Bias

The article is co-authored with Adam Glynn and Miguel Rueda. I am the lead author.

Adam Glynn and Miguel Rueda, on the one hand, and myself on the other hand had independently from each other conceived of the idea and had independently and separately from each other written a first draft. The current version is the result of joining our research.

The introduction was written by Glynn, Rueda, and me. The section "Understanding Conditional IV Identification Using Causal Graphs" was written by me. Specifically, the proposition was developed and proved by me. The sensitivity analysis was developed by Glynn, Rueda, and me. The corresponding parts in the paper were written by me. The implementation and data analysis were done by me. The corresponding parts in the paper were written by me. The conclusion was written by me.

Chapter 4: Causal Graphs for Survey Inference

The article is co-authored with Peter Selb. We share authorship. The idea was developed by both of us. Most of the paper was written by both of us. The introduction was written by Peter Selb. The section "Do adjustment variables need to correlate with survey selection and outcome?" as well as the Appendix was developed and written by me.