Analyzing Causal Mechanisms in Survey Experiments*

Avidit Acharya,† Matthew Blackwell,‡ and Maya Sen§

July 5, 2016

Abstract

We present an approach to investigating causal mechanisms in experiments that include mediators, in particular survey experiments that provide or withhold information as in vignettes or conjoint designs. We propose an experimental design that can identify the *controlled direct effect* of a treatment and also, in some cases, what we call an *intervention effect*. These quantities can be used in ways to address substantive questions about causal mechanisms, and can be estimated with simple estimators using standard statistical software. We illustrate the approach via two examples, one on characteristics of U.S. Supreme Court nominees and the other on public perceptions of the democratic peace.

*Comments and suggestions welcome. Many thanks to Paul Testa and to participants at the 2016 Midwest Political Science Association Conference for helpful feedback. Special thanks to Jessica Weeks and Mike Tomz for sharing their survey instrument with us.

†Assistant Professor of Political Science, Stanford University. email: avidit@stanford.edu, web: http://www.stanford.edu/~avidit.

‡Assistant Professor of Government, Harvard University. email: mblackwell@gov.harvard.edu, web: http://www.mattblackwell.org.

§Assistant Professor of Public Policy, Harvard University. email: maya_sen@hks.harvard.edu, web: http://scholar.harvard.edu/msen.

1 Introduction

In many experiments, such as survey experiments, subjects are exposed to different kinds of information to better understand how they form beliefs and opinions. For example, do survey respondents view black and white U.S. Presidential candidates differently? If so, do they continue to view the two candidates differently when provided with the additional cue that both are conservative Republicans? These research designs—which rely on vignettes, hypothetical examples, or the provision (or withholding) of information—are designed to test specific theories. Similarly, observational studies have also tried to exploit differences in mediating environments to gain better understanding of causal mechanisms. For the applied researcher, however, there are open questions. What are the quantities of interest being identified, and what do they mean? What is a suitable experimental design for investigating a causal mechanism?

In this paper, we present a general framework for understanding the quantities of interest identified by these sorts of research designs. Experiments that manipulate the information environment presented to subjects—including, for example, vignette or conjoint experiments—serve to identify the *controlled direct effect* of the treatment, which is the treatment effect net the effect of a mediator that is fixed at a particular value (Robins and Greenland, 1992). For example, using the illustration of the Presidential candidate experiment, presenting respondents with the information that both candidates are conservative Republicans and still seeing an effect associated with candidate race would be a controlled direct effect—that is, the effect of a racial cue net the effect of a partisan cue, which is held constant. From this type of design, it is possible to learn about these direct effects, but also causal interactions, which are differences in controlled direct effects.

We propose and analyze an experimental design that randomly varies the information provided to respondents with the purpose of exploring causal mechanisms. The decision to intervene on a specific attribute, such as party in the example above, can have consequences for the interpretation of the effects of other treatments in the design. To explore these choices and what they mean for our interpretation of

these experiments, we introduce a separate quantity of interest, which we call the *intervention effect*. This is the difference between setting just the treatment at some value (and letting the mediator take on whatever value it would take naturally) versus setting the treatment and mediator jointly at specific values. We show that the difference in two intervention effects can be interpreted as a combination of indirect (or mediated) effects and causal interactions. This decomposition is useful to applied researchers who want to learn if the mediator can explain the overall effect of a particular treatment.

Both quantities of interest can be (and, indeed, have been) applied in experimental settings, particularly in survey experiments. We demonstrate this using two illustrative examples. The first examines how the public evaluates candidates to the U.S. Supreme Court and shows that we can identify different quantities of interest depending on the information provided to the respondent. Specifically, showing the respondents information about candidate partisanship reduces the signal conveyed to the respondents by the candidate's race or ethnicity. It appears that most of the total effect of race can be explained by the inferred partisanship of the candidate. The second example replicates findings from Tomz and Weeks (2013) on the "democratic peace" theory showing that Americans are less likely to support preemptive strikes against democracies versus non-democracies. Using our framework, we are able to show that this difference is strengthened when information about potential threats are provided, suggesting that Americans are even more likely to support preemptive strikes against non-democracies when these countries are under threat.

This paper proceeds as follows. We first introduce the formalism and define the key terms of our inquiry. We also describe the simple motivating example that we use throughout, that of a survey experiment assessing how different characteristics influence public support for U.S. Supreme Court nominees. Next, we define our two main causal quantities of interest: (1) controlled direct effects and (2) intervention effects. We explain how these quantities apply not just to experiments (and survey experiments in particular, using our illustrative example), but also how they

can be used more broadly in observational contexts. We then analyze the two illustrative examples, both of which show that we can identify different quantities of interest depending on the information provided to the respondent. We conclude with a discussion of the additional implications of our study for applied research in political science.

2 Setting and Illustrative Example

We develop the main ideas using the simple example of a candidate choice survey experiment. Such experiments often examine how respondents react to changes in question wording vignettes, or information provided (for example using different profiles as would be done with a conjoint experiment). For example, suppose a researcher is interested in understanding how the public evaluates potential U.S. Supreme Court nominees (a topic explored in Sen, 2016). Suppose that the researcher is interested in understanding to what extent race and racial cues change the public's view of a potential nominee. An attractive design with which to explore this question would be one that presents respondents with two profiles: for example, one with a candidate simply identified to the respondents as African American and another with a candidate simply being identified as white. Hainmueller, Hopkins and Yamamoto (2013) illustrate a similar conjoint design via an example involving the U.S. Presidency.

Simply comparing the two profiles would allow for estimation of the treatment effect associated with the racial cue. However, without further information provided to the respondents, a simple design such as this one would fail to clarify the mechanism behind the treatment effect. For example, a negative treatment effect associated with the black racial cue could be attributed to straightforward racial animus. Or, a negative treatment effect among respondents could also be attributed to a prior belief that black nominees are more likely to be Democrats or left-leaning—not necessarily an unreasonable starting prior. Yet another possibility is that another treatment effect could be attributed to respondents thinking that

white candidates are more likely to come from lower tiers of the federal judiciary and are therefore more qualified. These three explanations point to very different substantive conclusions about the political environment and about minority nominees: the first would point to racial prejudice while the second and third use race as a heuristic for other characteristics.

If the researcher included information about the candidate's partisanship in her experiment, perhaps as part of the candidate profile, then she would be able to assess whether the second hypothesis is supported. If she included information about the candidate's professional background in the survey experiment, then she would be able to assess support for the third hypothesis. This kind of approach—common across political science—illustrates the reasoning for including more information in survey experiments. More broadly, the same kind of research design underlies many inquiries using vignettes, hypothetical examples, and manipulations of the information environment.

We view the goals of these types of experiments as twofold. First, and most obviously, researchers using these kinds of designs want to estimate the baseline causal effects of each attribute. Looking at our example again, is there an effect of nominee race on a respondent's choice? Second, given a particular total effect (or marginal component effect in the terminology of conjoint experiments) researchers want to understand *why* and *how* there is an effect. That is, we would like to know the mechanism by which the effect came to be—e.g., why does race affect a respondent's choice? The first of these questions is relatively straightforward in an experimental setting, and a large literature in statistics and political science has focused on the estimation of these treatment effects. The second question has been of increasing interest across social science—with most researchers looking at these questions proceeding in an ad hoc basis. Our goal here is to reason more formally about this second goal, that of investigating mechanisms. To do this, we turn now to explaining what we mean by a causal mechanism and how certain experimental designs facilitate their exploration.

2.1 Mechanisms, Mediation, and Interaction

A causal mechanism (1) provides an explanation for why and how a cause occurs—that is, what factors contributed to the causal effect that we see in front of us?—and, (2) in the spirit of counterfactual reasoning, explains what differences in terms of intervening or contextual forces would have produced a different result. To approach these questions, we start with some definitions, building from the framework introduced by VanderWeele (2015). We define a *causal mechanism* as either a description of (1) the causal process—that is, how a treatment affects an outcome—or (2) a causal interaction—that is, in what context does the treatment affect the outcome. Both causal processes and causal interactions speak to the mechanism by which a treatment affects an outcome, and both answer the questions we posed above. For that reason, both concepts give applied researchers insights that can be used to design better, more effectively-tailored, interventions.

Mechanisms as causal processes. The first of these, *mechanisms as causal processes*, describes how the the causal effect of a treatment might flow through another intermediate variable on causal pathway from treatment to outcome, or what is sometimes referred to as a causal pathway. The existence of a causal process—also called an indirect or mediated effect—tells us how the treatment effect depends on a particular pathway and gives us insight into how changes to the treatment—ones that might alter these pathways—would produce different treatment effects. In terms of our example looking at the U.S. Supreme Court, this might be how the race of the hypothetical candidate affects respondents' beliefs about the partisanship of the nominee, which in turn affects respondent choice.

Mechanisms as causal interaction. The second of these, *mechanisms as causal interactions*, describes how manipulating a secondary, possibly intermediate, variable can change the magnitude and direction of a causal effect. This is an important goal for many applied researchers: a causal interaction reveals how a treatment effect could be either altered or entirely removed through the act of intervening on

a mediating variable. In this sense, causal interactions speak to the context of a causal effect, as opposed to the pathway, and how altering this context can change the effectiveness of a particular intervention. In terms of our example involving hypothetical Supreme Court candidates, a straightforward example is partisanship. Providing respondents with information about a candidate's partisanship could substantially alter the effects associated with race, if, for example, race is a more salient consideration when the nominee is a copartisan.

It's worth emphasizing that causal interactions do not depend on the treatment causally affecting the mediator—which means that exploring mechanisms as causal interactions works well with experiments that randomly assign several attributes at once, such as conjoints or vignettes. For example, suppose a researcher randomly assigns respondents to see candidate profiles with different racial backgrounds and also with different partisan affiliations, i.e., with randomly assigned combinations. By design, race (the treatment) does not causally affect partisanship (the mediator) because both are randomly assigned. However, the effects of race on respondent evaluation of the hypothetical candidate may still nonetheless depend on what value partisanship (the mediator) is set to. And the interactions between the two, as we discussed above, yield insights into the mechanism by which race affects respondents' revaluations in situations where partisanship is not manipulated.

Other approaches. Other approaches exist in the literature. For example, in a recent paper on survey experiments, Dafoe, Zhang and Caughey (2016) refer to the changing nature of the treatment effects in the setting that we have in mind as "confounding." Under their framework, the true treatment effect of an randomized assignment is confounded by a respondents' beliefs over other features of the vignette driven by the experimental design.[1] The benefit of this approach is that it clarifies the connection between the experimental design and the beliefs of re-

[1] For example, using our illustration, if the researcher just provided respondents with information about the candidate's race, then any kind of treatment effect associated with race would be "confounded" by partisanship. That is, respondents might assume that candidates of certain racial or ethnic backgrounds have different partisanships.

spondents. Our approach differs in that we place no value-labeling on the various effects estimated with different designs. That is, we do not seek to estimate the "true" effect of some treatment, but rather we seek to understand *why* a particular treatment effect might exist.

Another approach is that of Imai, Tingley and Yamamoto (2013), who explore various experimental designs that help identify mediation effects. In many cases, these designs cannot point-identify these indirect effects, though bounds on the effects can be estimated from the data. However, these bounds may not even identify the direction of the effect. This highlights a limitation of some experimental designs in which it is impossible to unpack a causal mechanism in terms of processes and interactions. It also motivates our present set of questions—what can we learn or explain about a set of causal effects from these experimental designs?

3 Assumptions and quantities of interest

We now present the formalism. First, we consider a setting with two parallel survey experiments, which we indicate by $D_i = 0, 1$, where i is the subject. If both the treatment and the mediator are randomized for subject i, then $D_i = 0$; if only the treatment is randomized then $D_i = 1$. We call the first of these arms the *manipulated-mediator arm* and the second the *natural-mediator arm*. We denote the treatment in both experiments by $T_i \in \mathcal{T}$, where T_i can take on one of J_t values. To keep the discussion focused, we assume that there is only one attribute (such as race) in T_i, but below we discuss extending the framework to handle a multidimensional treatment, as in a conjoint design. There is also a potential mediator that takes value $M_i \in \mathcal{M}$ for subject i. The mediator can take one of J_m values. In our running example, $T_i = 0$ would, for example, indicate that the candidate was reported to be African American and $T_i = 1$ would indicate that the candidate was reported to be white. The mediator in this case would be partisanship, for example with $M_i = 0$ indicating that the candidate is a Democrat and $M_i = 1$ indicating that the candidate is a Republican.

To define the key quantities of interest, we rely on the usual potential outcomes framework for causal inference (Rubin, 1974; Holland, 1986; Neyman, 1923). In the natural-mediator arm, $D_i = 1$, we do not manipulate the mediator, so it has potential outcomes as a function of the treatment, $M_i(t)$, which is the value that the mediator would take for subject i if they were assigned to treatment condition t. For example, in our illustration, this would be the value that the mediator (partisanship) would take on for each respondent take if the respondents were given information only about candidate race.[2] In the manipulated-mediator arm, $D_i = 0$, both the treatment and the mediator would be assigned by the researcher. In our illustration this would mean providing respondents with race/ethnic information *and* partisan information about the hypothetical candidates, thus leaving respondents with no room with which to "set" a value of the mediator.

In either experiment, each subject has a potential outcome associated with every combination of the two factors, $Y_i(t, m, d)$, which is the value that the outcome would take if T_i, M_i and D_i were set to values t, m, and d, respectively. We only observe one of these possible potential outcomes, $Y_i = Y_i(T_i, M_i, D_i)$, which is potential outcome evaluated at the observed combination of the treatment and the mediator. As in Imai, Tingley and Yamamoto (2013), we make the following consistency assumption: for all $(t, m) \in \mathcal{T} \times \mathcal{M}$, if $M_i(t) = m$, then

$$Y_i(t, M_i(t), 0) = Y_i(t, m, 1).$$

The assumption states that the value of the outcome should be the same in instances where the mediator takes on the same value either from direct manipulation or through the natural causal process. In our running example, this would mean that a respondent's support for the candidate is the same regardless of whether the respondent (let's say) infers that the candidate is a Democrat from the racial information as opposed to whether he or she was actually provided with the explicit cue

[2] In this case, perhaps many respondents would assume that a candidate identified as black is Democrat. Such a presumption would be in line with what Dafoe, Zhang and Caughey refer to as confounding.

that the candidate is a Democrat. This assumption may be problematic if manipulating the mediator has consequences for other variables of interest.

The consistency assumption enables us to write the potential outcomes simply as $Y_i(t, m) = Y_i(t, m, d)$. In the natural-mediator arm, with $D_i = 1$, the mediator takes its natural value—that is, the value it would take under the assigned treatment condition. We sometimes write $Y_i(t) = Y_i(t, M_i(t))$ to be the potential outcome just setting the value of the treatment. We also make a more general consistency assumption that connects the observed outcomes to the potential outcomes, such that $Y_i = Y_i(T_i, M_i)$ and $M_i = M_i(T_i)$.

We make a few randomization assumptions that follow directly from the design. We assume that in the natural-mediator arm, the treatment is randomized and, in the manipulated-mediator arm, that both the treatment and the mediator are randomized. We also assume that the assignment to these arms is also random. Thus, for all $(t, t', m) \in \mathcal{T} \times \mathcal{T} \times \mathcal{M}$,

$$\begin{aligned} \{Y_i(t, m), M_i(t')\} &\perp\!\!\!\perp D_i \\ \{Y_i(t, m), M_i(t')\} &\perp\!\!\!\perp T_i | D_i = 0 \\ Y_i(t, m) &\perp\!\!\!\perp \{T_i, M_i\} | D_i = 1. \end{aligned}$$

To extend this analysis to observational data, these assumptions can be generalized to accommodate covariates, both pretreatment and intermediate (Acharya, Blackwell and Sen, 2016).

3.1 Quantities of interest: indirect, interaction, and intervention effects

Causal effects are the differences between potential outcomes. For example, the individual (total) causal effect of treatment can be written as:

$$TE_i(t_a, t_b) = Y_i(t_a) - Y_i(t_b) = Y_i(t_a, M_i(t_a)) - Y_i(t_b, M_i(t_b)), \tag{1}$$

where t_a and t_b are two levels in $\mathcal{T}$. As is well-known, however, individual-level effects like these are difficult to estimate without strong assumptions because we only

observe one of the J_t potential outcomes for any particular unit i. Given this, most investigations of causal effects focus on average effects. For example, the *average treatment effect* (ATE) is the difference between the average outcome if the entire population were set to t_a versus the average outcome if the entire population were set to t_b. We write this as $TE(t_a, t_b) = \mathbb{E}[TE_i(t_a, t_b)] = \mathbb{E}[Y_i(t_a) - Y_i(t_b)]$, where $\mathbb{E}[\cdot]$ is the expectation operator defined over the joint distribution of the data.

Controlled Direct Effects. The manipulated-mediator arm allows us to analyze the joint effect of both of these interventions. In particular, we can define the individual-level *controlled direct effect* as the effect of treatment for a fixed value of the mediator:

$$CDE_i(t_a, t_b, m) = Y_i(t_a, m) - Y_i(t_b, m). \tag{2}$$

Referring back to our illustrative example involving the U.S. Supreme Court, the total treatment effect is the difference in support for a hypothetical black candidate versus a hypothetical white candidate for unit i. The controlled direct effect, on the other hand, would be the difference in support between these two candidates where it is additionally revealed that the two candidates are of the same political party. Of course, as with the total effect, one of the two potential outcomes in the CDE_i is unobserved so we typically seek to estimate the *average controlled direct effect* (ACDE), which is $CDE(t_a, t_b, m) = \mathbb{E}[CDE_i(t_a, t_b, m)] = \mathbb{E}[Y_i(t_a, m) - Y_i(t_b, m)]$.

Natural Indirect Effects. The *natural indirect effect* of the treatment through the mediator is:

$$NIE_i(t_a, t_b) = Y_i(t_a, M_i(t_a)) - Y_i(t_a, M_i(t_b)). \tag{3}$$

This is the effect of changing the mediator with a change in treatment, but keeping treatment fixed at a particular quantity. As the name implies it represents an indirect effect of treatment through the mediator. This quantity will be equal to zero if either (1) the treatment has no effect on the mediator so that $M_i(t_a) = M_i(t_b)$, or (2) the mediator has no effect on the outcome. Both of these conditions are

intuitive given the usual motivation of indirect effects as multiplicative: the effect of treatment on the mediator is multiplied by the the effect of the mediator on the outcome.[3] As above, we define the *average natural indirect effect* (ANIE) to be $NIE(t_a, t_b) = \mathbb{E}[NIE_i(t_a, t_b)] = \mathbb{E}[Y_i(t_a, M_i(t_a)) - Y_i(t_a, M_i(t_b))]$.

Reference Interactions. To capture the interaction between T_i and M_i, we introduce the so-called *reference interaction*, which is the difference in controlled direct effects between the reference category m and the natural value of the mediator under t_b, or $M_i(t_b)$:

$$RI_i(t_a, t_b, m) = \sum_{\tilde{m} \in \mathcal{M} \setminus \{m\}} \mathbb{I}\{M_i(t_b) = \tilde{m}\} \left[CDE_i(t_a, t_b, \tilde{m}) - CDE_i(t_a, t_b, m)\right] \quad (4)$$

This would be the difference in the CDE of black versus white nominees between the inferred partisanship under a white nominee and the manipulated partisanship. When we average this quantity over the population, we end up with a summary measure of the amount of interaction between the treatment and mediator:

$$\begin{aligned} RI(t_a, t_b, m) &= \mathbb{E}[RI_i(t_a, t_b, m)] \\ &= \sum_{\tilde{m} \in \mathcal{M} \setminus \{m\}} \mathbb{E}\big[CDE_i(t_a, t_b, \tilde{m}) - CDE_i(t_a, t_b, m) | M_i(t_b) = \tilde{m}\big] \mathbb{P}[M_i(t_b) = \tilde{m}] \end{aligned} \quad (5)$$

This quantity, which we call the average reference interaction effect (ARIE) is the average interaction we see in the controlled direct effect using $M_i = m$ as a reference category (VanderWeele, 2015, p. 607). It will be equal to zero when either (1) there is no treatment-mediator interaction for this particular CDE, or (2) there is zero probability of the nature value of the mediator under t_b being equal to anything other than m. In both cases there is no interaction, either because the treatment effects or the natural value of the mediator doesn't vary. It is possible for this quantity to be zero if there is exact cancellations in the interactions across the population, but this is both rare and dependent on the baseline category, m.

[3] Even with heterogeneous treatment effects or a nonlinear model, the NIE provides a useful heuristic at the individual level.

Intervention Effects. We introduce a new quantity based on the changes induced by intervening on the mediator. The *intervention effect* is the difference in the outcomes resulting from allowing the mediator to take on a value associated with a particular treatment value versus the outcome resulting from fixing both the mediator and the treatment:

$$IE_i(t, m) = Y_i(t) - Y_i(t, m) = Y_i(t, M_i(t)) - Y_i(t, m) \tag{6}$$

This quantity is often of interest for applied researchers. A fruitful analogy here would be a study looking at the effects on weight gain of two prescriptions: diet and exercise. Intervention effects would be appropriate if a researcher was interested in giving people a joint prescription of diet and exercise rather than just telling people to exercise, which could cause people to eat more. Specifically, in this case, she would be interested in knowing the effect of the additional dietary prescription. Using our illustration of candidates to the U.S. Supreme Court, the intervention effect would be the difference in outcomes between (1) just showing respondents that a hypothetical candidate is black (for example) and (2) showing respondents that a hypothetical candidate is black *and* a Democrat. The difference between support in these conditions might be thought of a measure of the support for a black nominee that is lost when it is revealed that the nominee is a Democrat.

The *average intervention effect* (AIE) is $IE(t, m) = \mathbb{E}[IE_i(t, m)] = \mathbb{E}[Y_i(t) - Y_i(t, m)]$. This is a suitable quantity of interest for experiments that provide additional information to some, but not all respondents. This may be the case in conjoint experiments, vignette experiments, or certain field experiments where the on-the-ground intervention concerns manipulation of the informational environment. The AIE is distinct from either an indirect effect or the direct effect of the mediator on the outcome. The indirect effect, as usually given in the causal inference literature, compares the potential outcome under one treatment, $Y_i(t_a, M_i(t_a))$ to the potential outcome when the mediator is set to its unit-specific value under the other treatment, $Y_i(t_a, M_i(t_b))$. In practice, this second quantity is impossible to observe without further assumptions because it requires simultaneously observing a unit under t_a (for the outcome) and t_b (for the mediator). Since

we never observe both of these states at once, any inference involving that quantity will require strong and perhaps unrealistic assumptions. The direct effect of the mediator in this case is the effect of changing the value of the mediator (in the second experiment): $Y_i(t, m_c) - Y_i(t, m_d)$. This differs from the intervention effect because there will be some units for whom $M_i(t) = m$ and so the intervention effect is 0 for those units. In our example, this is best illustrated by the fact that some respondents are unaffected by the revelation that the hypothetical candidate is a Democrat because they would have already assumed she was a Democrat based on her race. Thus, in the binary mediator case, the average intervention effect is a mixture of zero effects for some units and the direct effect of the mediator for other units.

Difference in Intervention Effects. Finally, the *difference in intervention effects* (ΔIE) is

$$\begin{aligned} \Delta_i(t_a, t_b, m) &= IE_i(t_a, m) - IE_i(t_b, m) \\ &= [Y_i(t_a) - Y_i(t_a, m)] - [Y_i(t_b) - Y_i(t_b, m)] \,. \end{aligned} \tag{7}$$

This quantity tells us how the intervention effect varies by level of the treatment and is equivalent to the difference between the treatment effects in the two experiments:

$$[Y_i(t_a, M_i(t_a)) - Y_i(t_b, M_i(t_b))] - [Y_i(t_a, m) - Y_i(t_b, m)]$$

In the context of the Supreme Court nominee experiment, this would tell us the relative change in support for black nominees versus white nominees when each is revealed to be a Democrat. Alternatively, it would be the difference between the total effect of a black versus white nominee and the controlled direct effect for the same difference when both candidates are Democrats. Finally, the *average difference in intervention effects* (AΔIE) is $\Delta(t_a, t_b, m) = \mathbb{E}[IE_i(t_a, m) - IE_i(t_b, m)] = IE(t_a, m) - IE(t_b, m)$, which is simply the difference in average intervention effects at two levels of the treatment given the manipulated level of the mediator m.

3.2 How Intervention Effects Help us Understand Causal Mechanisms

In this section, we explain what the difference in two intervention effects can teach us about the underlying causal mechanisms. Through rearrangement of the terms in (7), it easy to see that

$$\Delta_i(t_a, t_b, m) = \underbrace{TE_i(t_a, t_b)}_{\text{treatment effect}} - \underbrace{CDE_i(t_a, t_b, m)}_{\text{controlled direct effect}} \tag{8}$$

which implies that the average difference in intervention effects is also the difference between the ATE in the natural-mediator arm and the ACDE in the manipulated mediator arm. Thus, we can interpret the difference as the amount of the overall treatment affect that remains after intervening on the mediator.

Under consistency, we can also characterize the difference in intervention effects using the following decomposition:

$$\Delta_i(t_a, t_b, m) = \underbrace{NIE_i(t_a, t_b)}_{\text{indirect effect}} + \underbrace{RI_i(t_a, t_b, m)}_{\text{interaction effect}} \tag{9}$$

The residual effect of treatment after intervening on the mediator, then, is a combination of an indirect effect of treatment through the mediator and an interaction effect between the treatment at the mediator. This quantity, then, is a combination of the two aspects of a causal mechanism: the causal process, represented by the indirect effect, and the causal interaction, represented by the interaction effect. Thus, we can interpret Δ as the portion of the ATE that can be explain by M_i, either through indirect effects or interactions.

In the candidate choice experiment, the difference in intervention effects is the combination of two separate components. The first is the indirect effect of race on choice through party. The second is the interaction between party and race, for those units that would think a white candidate (t_b) is a Republican, $M_i(t_b) = 1$, scaled by the size of this group. This second component will be close to zero when the interaction effect is 0 or when party and race are tightly coupled so that very

few people imagine than a white candidate is a Republican. In some contexts, then, it may be plausible that $\mathbb{P}[M_i(t_b) = 1] \approx 0$ and the difference in the intervention effects can be interpreted as, essentially, the indirect effect. For example, given that so few African Americans identify with the Republican party, it might be reasonable to assume that almost all respondents would infer that such a nominee is a Democrat. Even when these conditions do not hold, the difference in intervention effects still has an interpretation as being a combination of the indirect effect and an interaction between the treatment and the mediator.

Under the above assumptions, it is not possible to tease apart the relative contribution of the indirect and interaction effects in contributing to the difference in intervention effects. In order to do so, we require strong assumptions at the individual level that either assume away any interactions at the individual level or assume that the natural value of the mediator is unrelated to the interaction effects (Imai, Keele and Yamamoto, 2010). If, for instance, we assume that the CDE does not vary with m at the individual level then $CDE_i(t_a, t_b, m_c) - CDE_i(t_a, t_b, m_d) = 0$ and the difference in intervention effects is exactly equal to the indirect effect. This approach is problematic because such "no interaction" assumptions are highly unrealistic in most settings. Furthermore, this approach is only required when one takes the strong position that causal mechanisms only relate to causal processes and indirect effects. When one takes a broader view of causal mechanisms as we do, it becomes acceptable to allow for indirect effects and interactions to contribute to a measure of a causal mechanism.

3.3 Extension to Conjoint Experiments

This basic setup can be easily extended to conjoint experiments where there are several attributes being manipulated at once and several separate profiles being shown to each respondent. This would mean that T_i is actually a multidimensional vector indicating the set of profiles provided to respondent i to rate. For example, our treatment might include information about the race of the proposed nominee, but it also might include information about the religion, age, and educational

background of the nominee. In this setting, Hainmueller, Hopkins and Yamamoto (2013) have shown that, under the assumption of no-profile order effects and no carryover effects, simple difference-in-means estimators that aggregate across respondents and rating tasks is unbiased for what they call the *average marginal component effect* or AMCE. This quantity is the marginal effect of one component of a profile, averaging over the randomization distribution of the other components of the treatment—the effect of race, averaging over the distribution of religion, age, and educational background, for instance. In many ways, this quantity is very similar to the ATE in the above discussion, with the caveat that the randomization distribution in the conjoint experiment might differ in fundamental ways from the distribution of beliefs without intervention. This allows us to think of the difference in intervention effects in this setting as both how the AMCE responds to additional intervention in the profile, but also as a measure of how the additional intervention in the profile helps explain the "total" effect of the AMCE.

3.4 Relevance for Observational Studies

Here we digress from our focus on experimental designs by relating the approach to the approach taken by Acharya, Blackwell and Sen (2016) for observational studies. Typically, it is useful to think of observational studies as also having experimental interpretations; in this sense, we show that the difference in intervention effects also has a conceptual meaning in observational studies.

Because the difference in intervention effects is also the difference in treatment effect in the natural-mediator arm and the manipulated mediator arm (see equation (8)), it is the case that when the average treatment effect and the controlled direct effect are both identified in an observational study, the difference in intervention effects is also identified. And, when this is the case, the same estimator that Acharya, Blackwell and Sen (2016) use to estimate the controlled direct effect can be used in estimating the difference in intervention effects (by subtracting the estimate for the CDE from the estimate for the treatment effect).

The fact that the treatment effect can be decomposed into a controlled direct

effect and difference in intervention effect (rearrange equation (8)) suggests that the difference in intervention effects has a conceptual meaning in observational studies even though, in practice, directly intervening on the mediator is typically impossible in an observational study. For example, Acharya, Blackwell and Sen (2016) considered an example from Alesina, Giuliano and Nunn (2013) who claim that historical plough use affects contemporary attitudes towards women, and attempted to rule out the possibility that the effect works through contemporary mediators such as income. Taking contemporary income as the potential mediator in the effect of historical plough use on contemporary attitudes towards women, the difference in intervention effects in this example is the following. First consider intervening on a unit where income is set to a pre-specified level, and varying the level of plough use from the realized level to another pre-specified level. Then consider performing the same intervention in an otherwise identical unit with a different level of plough use. The difference in intervention effects is the difference in effects between these two cases. If the two intervention effects are the same, we might interpret this as evidence that contemporary income does not "explain" the effect of historical plough use. However, if they are different, we might interpret it as evidence that it does explain some (or all) of it.

4 Estimation

We now turn to identification and estimation strategies. Under the assumptions above, it is straightforward to show that the difference in intervention effects is identified as:

$$\begin{aligned}\Delta(t_a, t_b, m) = &\big[\mathbb{E}[Y|T_i = t_a, D_i = 1] - \mathbb{E}[Y|T_i = t_a, M_i = m, D_i = 0]\big] \\ &- \big[\mathbb{E}[Y|T_i = t_b, D_i = 1] - \mathbb{E}[Y|T_i = t_b, M_i = m, D_i = 0]\big]\end{aligned} \tag{10}$$

We omit the proof given that it is a straightforward application of standard results in experimental design.

How might we estimate this quantity with our experimental samples? A simple plug-in estimator would replace the expectations above with their sample counter-

parts. For instance, we would estimate $\mathbb{E}[Y_i|T_i = t_a, D_i = 1]$ with:

$$\widehat{\mathbb{E}}[Y_i|T_i = t_a, D_i = 1] = \frac{\sum_{i=1}^{N} Y_i \mathbb{I}\{T_i = t_a, D_i = 1\}}{\sum_{i=1}^{N} \mathbb{I}\{T_i = t_a, D_i = 1\}} \tag{11}$$

Replacing each of the expectations in (10) in a similar fashion would produce an unbiased estimator for Δ. A convenient way to produce this estimator is through linear regression on a subset of the data. Specifically, to estimate these quantities, it is sufficient to subset the manipulated-mediator arm arm ($D_i = 0$) to those who have $M_i = m$ and regress Y_i on an intercept, a vector of $J_t - 1$ dummy variables for the levels of T_i, W_{it}, the experimental arm dummy, D_i, and interactions $W_{it}D_i$. Under this regression model, if t_b is the omitted category, then the coefficient on W_{it_a} is an unbiased estimator of $ATE(t_a, t_b)$ and the coefficient on $W_{it_a}D_i$ will be equivalent to the above nonparametric estimator for $\Delta(t_a, t_b, m)$. Note that because this regression model is fully saturated, it makes no assumptions about the functional form of the conditional expectation of Y_i and is equivalent to an estimator that estimates effects within all strata of the T_i and D_i. One benefit of this approach is that it is not necessary to measure M_i in the natural-mediator arm, $D_i = 0$.

Estimation with conjoint experiments under complete randomization across and within experimental arms is straightforward. Let T_{ikl} represents the lth attribute of the kth profile being evaluated, which can take on J_l possible values, and let Y_{ik} is subject i's response to the kth profile. Hainmueller, Hopkins and Yamamoto (2013) showed that it is possible to estimate the ACME by regressing Y_{ik} on the $J_l - 1$ dummy variables for the attribute of interest. The coefficients on each dummy variable in this case would be unbiased estimates of the ACME of that treatment level relative to the baseline group. To estimate the difference in intervention effects for a particular attribute, we simply interact these dummy variables with the experimental-arm indicator, D_i. With multiple rating tasks per respondent, there is within-respondent clustering and so variance estimation should be done either with cluster-robust standard errors or with a block bootstrap, where respondents are resampled with replacement.

5 Experimental Analysis of Direct Effects and Mechanisms

5.1 Study #1: Conjoint Experiment Involving Candidates to the U.S. Supreme Court

As a first application of these ideas, we take an example from Sen (2016) on how the public views nominations to the U.S. Supreme Court. The example provides an attractive illustration for the reason that, because many view the court as an apolitical institution, the ideology of Supreme Court candidates is often noisily conveyed to the public. However, as Sen (2016) shows, partisan and ideological cues are among the single most important signal that respondents evaluate in assessing potential nominees to the high court—significantly more important than race or gender identity. For that reason, the data are unique in the sense that half of the respondents in the study were randomly assigned to see a conjoint profiles that contained partisan information about a potential candidate and half were assigned to see profiles that contained no such information.

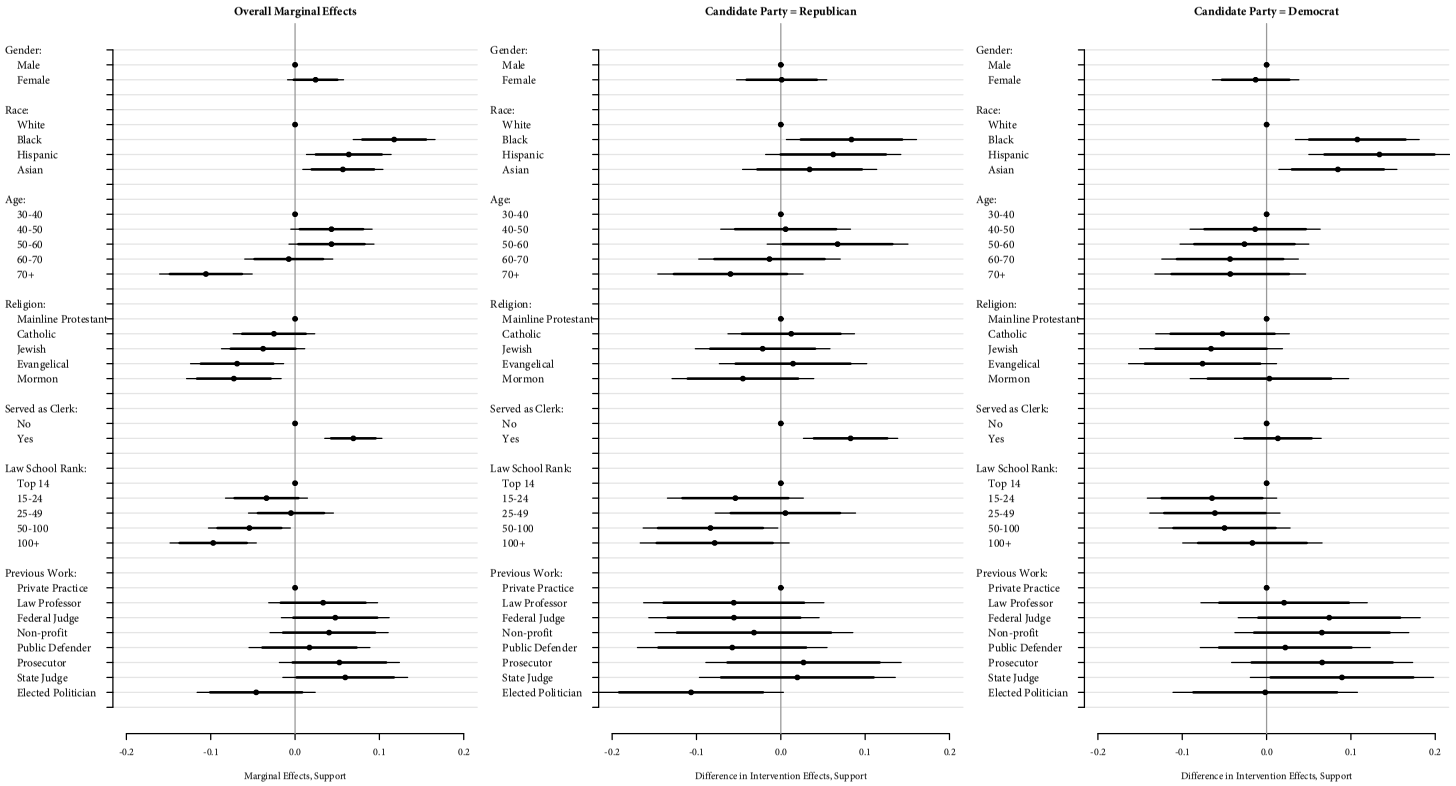
This experimental design matches our setting well and allows us to explore the implications of our framework. For example, in the absence of partisan cues, racial information contained in the profiles may activate in respondents presuppositions about partisan leanings. It would be logical for respondents to place strong priors on a potential candidate identified as black as being liberal or liberal leaning (compared to candidates identified by the profile as being white). Thus, the AMCE of the "black" racial cue should be positive for Democratic respondents. However, introducing a mediator such as partisanship, as was done for half of the respondents, allows us to estimate perhaps a more appealing and substantively meaningful quantity of interest, the controlled direct effect of the "black" racial cue. From these two experimental arms, we can estimate the difference in intervention effects, $\Delta(t_a, t_b, m)$. If this quantity was also positive for Democratic respondents, it would indicate that some portion of the positive AMCE of race is due to inferred

partisanship, either through indirect effects or interactions. Of course, we can estimate the AMCE and the difference in intervention effect for each attribute of the conjoint profiles.

The experimental design embedded two conjoint experiments, to which respondents were randomly assigned. In one arm of the experiment, respondents rated profiles that included race, gender, age, religion, previous work experience, and law school rank, but omitted any information about the partisanship of the nominee. In the other arm, the profiles included information about the party of the nominee in addition to all of the attributes. We focus on the respondents who identify as Democrats for the sake of exposition. This way, copartisanship between the respondent and the profile can be viewed as randomly assigned in the partisanship arm of the experiment. To analyze this experiment, we estimate the total AMCEs from the natural-mediator arm, and then estimate the difference in intervention effects when the mediator, here partisanship of the nominee, is set to Republican (non-copartisan) and Democrat (copartisan). Given the above discussion, these differences in intervention effects will give us some sense of whether or not partisanship participates in a mechanism for the various marginal effects of each component.

Figure 5.1 shows the results from this study, with the total AMCEs in the left panel, the difference-in-intervention effects for Republican profiles and Democratic profiles in the middle and right panels, respectively. The Figure shows both 90% and 80% confidence intervals for each point estimate, based on cluster-robust standard errors. From the total effects, we can see that Democratic respondents are more like to support minority nominees, nominees that served as a clerk, nominees that attended higher-ranked law schools, and nominees who are younger than 70. But these effects are in the condition where respondents had no access to information about the partisanship of the nominee.

Is it possible that some of these effects are due to different inferred partisanship of the nominee from these characteristics? The differences-in-intervention effects tell us this exactly. For instance, the positive difference-in-intervention effects for


Overall Marginal Effects
Candidate Party = Republican
Candidate Party = Democrat
Gender:
Male
Female
Race:
White
Black
Hispanic
Asian
Age:
30-40
40-50
50-60
60-70
70+
Religion:
Mainline Protestant
Catholic
Jewish
Evangelical
Mormon
Served as Clerk:
No
Yes
Law School Rank:
Top 14
15-24
25-49
50-100
100+
Previous Work:
Private Practice
Law Professor
Federal Judge
Non-profit
Public Defender
Prosecutor
State Judge
Elected Politician
-0.2
-0.1
0.0
0.1
0.2
Marginal Effects, Support
Difference in Intervention Effects, Support
Difference in Intervention Effects, Support

the racial minority effects are generally positive, meaning that it appears that partisanship does play a part in the causal mechanism for these attributes. These differences are especially acute for the effect of a black nominee versus a white nominee, which makes sense given the empirical distribution of partisanship across these racial groups. These differences imply that there is either an indirect effect of race on support through inferred partisanship or that there is a positive interaction between race and partisanship. Even though we cannot differentiate between these two sources of partisanship as a causal mechanism, it appears that partisanship does offer an explanation for the overall AMCE of race that we in the natural-mediation arm. For the other attributes, there appears to be less evidence that partisanship matters, which makes sense given that the racial information likely gives the strongest partisan cues of these attributes.

5.2 Study #2: Public Opinion and Democratic Peace

As a second application, we replicate the experimental study of Tomz and Weeks (2013), which explored whether American respondents are more likely to support preemptive military strikes on non-democracies versus democracies. To examine this, Tomz and Weeks present respondents with different country profiles and ask respondents whether they would, or would not, support preemptive American military strikes. They randomly assigned various characteristics of these profiles, including (1) whether the country is a democracy, (2) whether the country had a military alliance with the U.S., and (3) whether the country had a high level of trade with the U.S. Of particular interest to us here is that, leveraging a follow-up question, the authors use a mediation analysis to explore how *perceptions of threat* may mediate the effect of democracy on support for a strike. However, their mediation analysis requires that there be no unmeasured confounders between perceptions of threat and support for an attack, perhaps an unreasonably strong assumption.

In our replication, which we fielded using an Mechanical Turk sample of 1,247

respondents,[4] we added a second manipulation arm to this experiment that allows us assess whether perceptions of threat may play a role in explaining the overall effect of democracy without this problematic assumption. Specifically, following the original experimental design, we randomly assigned different features of the country in the vignette using the same criteria as Tomz and Weeks. We then manipulate one additional treatment condition. Some respondents were given the experimental design exactly as it was in Tomz and Weeks (2013), with no information given about the threat that the hypothetical country poses. In the manipulated mediator arm, on the other hand, the vignette provides the following additional information about the threat: "The country has stated that it is seeking nuclear weapons to aid in a conflict with another country in the region."[5] Note that it is still possible to identify and estimate $\Delta(t_a, t_b, m)$ even when there is only one value of M_i in the manipulated mediator arm, as is the case here.

Figure 5.2 shows the results from this replication. The analysis shows that, first, we are mostly able replicate Tomz and Weeks's finding that respondents are less likely to support a preemptive strike against a democracy versus a non-democracy (bottom-most coefficient, which is negative). However, this difference is not statistically significant, which is understandable since the number of units used to estimate the ATE here is roughly half the number used in the original experiment. Second, the ACDE of democracy with the information about threat held constant (at a high level) is more than double in magnitude than the ATE and statistically significant—an unusual instance in the sense that the ACDE is actually larger in magnitude than the ATE. Tomz and Weeks (2013) found a negative indirect effect of democracy through potential threat, which would imply that there should be either no or a positive natural direct effect of democracy on support for a preemptive strike. Here we find the opposite—with a potential threat revealed, democracy has an even stronger negative effect on support for a strike.

[4] The experiment was fielded online in June of 2016. The entire survey took around 5 minutes. The MTurk sample was restricted to adults aged 18 or older residing in the United States.

[5] The language for this manipulation comes from the measured mediator from the original study where Tomz and Weeks (2013) found a large effect of democracy on respondents' perceptions that the country would threaten to attack another country.

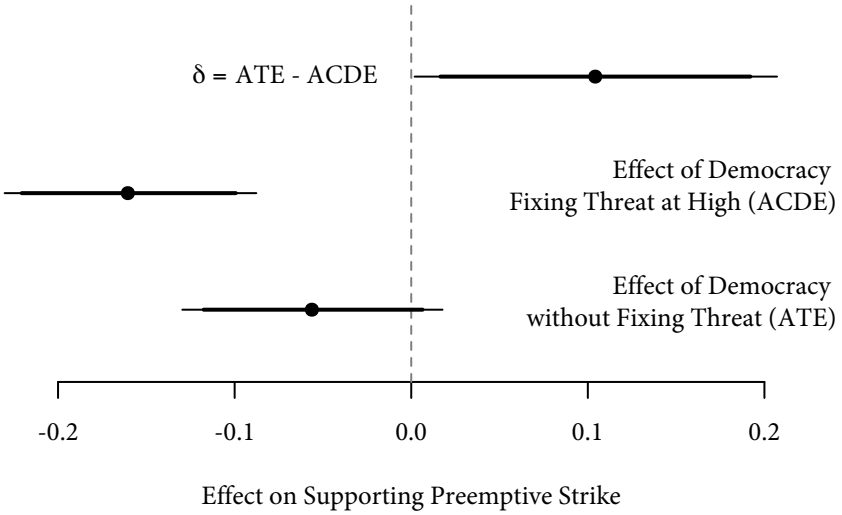


Figure 1: Results from the replication of Tomz and Weeks (2013). Data from a Mechanical Turk survey experiment ($N = 1247$).

There are two ways to reconcile these findings. First, if the difference between the ATE and the ACDE, Δ, is a combination of the indirect effect and an interaction effect, then it could be the case that there is a large, positive interaction between democracy and threat. That is, the effect of democracy might be larger when the profiled country seeks to use their weapons against an enemy, since perhaps respondents trust that democracies are more likely to be on the "right side" as compared to non-democracies. Another possible explanation for the differences in effects is that the negative indirect effect in Tomz and Weeks (2013) was biased due to a violation of the sequential ignorability assumption. This could occur if, for instance, the democratic status of the country affected an overall impression of the country, which then affected both support for a strike and perceptions of threat. Of course, these two explanations are not mutually exclusive and could work together to produce the large, positive Δ we see in this study.

6 Conclusion

We conclude by providing an assessment of how our framework may be useful for applied researchers. The most interesting political science questions focuses on when and how effects operate. Within the context of survey experiments, moreover, additional efforts have gone toward manipulating different components of information in order to tease apart causal mechanisms. The quantities of interest that we discuss here—controlled direct effects, intervention effects, and differences in intervention effects—speak directly to these questions.

How can applied researchers best leverage these quantities of interest? First, applied researchers need to give careful thought as to which quantity of interest best suits their needs. The controlled direct effect is particularly useful in instances where applied researchers need to "rule out" that a potential narrative is driving their results. For example, in our illustration of the U.S. Supreme Court, a plausible research inquiry is that the researcher in question needs to rule out the counter-argument that different priors about partisanship are driving her findings regarding the treatment effect of race.[6] On the other hand, the intervention effect is perhaps a more intuitive step, as it represents the difference associated with intervening on mediator as opposed to allowing the mediator to take on its "natural" value. In this sense, examining intervention effects is best used by applied researchers trying to understand the effect of a mediator on outcomes in a "real world" context. This may be of particular concern to those researchers particularly keen on emphasizing the external validity of experimental findings. Finally, the difference in intervention effects is a quantity that measures the extent to which the overall ATE of the treatment can be explained by the mediator. This quantity is a combination of an indirect effect and an interaction effect, both of which we interpret as being measures of how the mediator participates in a causal mechanism.

Assessing which of these quantities of interest suit applied researchers' needs is the first step. The second is estimation. We provided a number of different ways

[6] More on this sort of strategy, and also on the appropriateness of including post-treatment covariates, in the observational context is found in Acharya, Blackwell and Sen (2016).

in which these two quantities of interest can be estimated. In the survey context, providing respondents with different levels of information (that is, manipulating or fixing the treatments and mediators) in various ways will easily identify one or both quantities of interest. We also note that survey experiments, and conjoint experiments in particular, perhaps have the most flexibility in randomizing potential mediators. Thus, as our examples show, survey experiments enable the straightforward identification of both controlled direct effect and intervention effects—making them particularly flexible for applied researchers.

Lastly, the situation is more complicated in an observational context. However, oftentimes intervention effects may be identified in a straightforward fashion via natural experiments—that is, naturally occurring instances were the mediator is fixed for some groups but not for others, and the researcher is interested in the effect of the additional "information" or naturally occurring intervention. Future research might address additional quantities that may interest applied researchers an observational setting.

Bibliography

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* . In Press.
URL: http://www.mattblackwell.org/files/papers/direct-effects.pdf

Alesina, Alberto, Paola Giuliano and Nathan Nunn. 2013. "On the origins of gender roles: Women and the plough." *Quarterly Journal of Economics* 128(2):469–530.

Dafoe, Allan, Baobao Zhang and Devin Caughey. 2016. "Confounding in Survey Experiments: Diagnostics and Solutions." Working paper.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2013. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis (Winter* 22(1):1–30.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81(396):945–960.
URL: http://www.jstor.org/stable/2289064

Imai, Kosuke, Dustin Tingley and Teppei Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 176(1):5–51.

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):51–71.
URL: http://projecteuclid.org/euclid.ss/1280841733

Neyman, Jerzy. 1923. "On the application of probability theory to agricultural experiments. Essay on Principles. Section 9." *Statistical Science* 5:465–480. Translated in 1990, with discussion.

Robins, James M. and Sander Greenland. 1992. "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology* 3(2):143–155.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 6:688–701.

Sen, Maya. 2016. "How Political Signals Affect Public Support for Judicial Nominations: Evidence from a Conjoint Experiment." Working Paper.
URL: http://scholar.harvard.edu/files/msen/files/conjoint-judicial-nominations.pdf

Tomz, Michael R. and Jessica L. P. Weeks. 2013. "Public opinion and the democratic peace." *American Political Science Review* 107(04):849–865.

VanderWeele, Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.