# Analyzing Causal Mechanisms in Survey Experiments*

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

March 30, 2017

**Abstract**

We present an approach to investigating causal mechanisms in survey experiments that leverages the provision or withholding of information on mediating variables. These designs can identify the overall average treatment effect, the controlled direct effect of a treatment, and the difference between these two effects. They can also provide evidence for a broader understanding of causal mechanisms that encompasses both indirect effects and interactions, and they are also valid under weaker assumptions than those needed for mediation analysis. In addition, the main quantities of interest can be estimated with simple estimators using standard statistical software. We further show how these designs can speak to causal mechanisms even when only imperfect manipulation of the mediating variable is available. These methods help clarify key design choices in a broad class of increasingly popular experiments, including survey vignettes and conjoint designs. We illustrate these approaches via two examples, one on evaluations of potential U.S. Supreme Court nominees and the other on public perceptions of the democratic peace.

*Comments and suggestions welcome. Many thanks to Josh Kertzer, Paul Testa, Teppei Yamamoto, 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

Survey experiments have been used in the social sciences to detect whether causal effects exist. Understanding *why* a causal effect exists has, however, also become an important goal for researchers. To address this issue, survey experiments often manipulate the information given to respondents with the aim of shedding light on causal mechanisms. For example, do survey participants view otherwise identical profiles of black and white U.S. Presidential candidates differently? Would opinions change if people were provided with the additional cue that both candidates are Republicans? Does a respondent's propensity to support a preemptive strike by the U.S. against a nuclear power depend on whether the target country is a democracy? And, do any differences between democracies and non-democracies persist if we also tell respondents that the country poses little threat to the U.S.? As these questions illustrate, manipulating the informational environment of a survey experiment can reveal substantively important patterns.

In this paper, we show how the choice to intervene on a specific piece of information in a vignette or conjoint survey experiment can change the quantity of interest identified in substantively meaningful ways, giving insights into causal mechanisms. First, we review the consequences of intervening on one attribute (the mediator) when exploring the mechanisms of another attribute (the treatment). Experimental designs that use such interventions, proposed in the context of mediation analysis by Imai, Tingley and Yamamoto (2013), can identify the overall average treatment effect and the *controlled direct effect* of an attribute. This latter quantity is the treatment effect with another (potentially mediating) attribute held fixed at a particular value (Robins and Greenland, 1992). For example, in the Presidential candidate experiment, presenting respondents with the information that both candidates are Republicans and still seeing an effect associated with candidate race would be a controlled direct effect—that is, the effect of a racial cue with partisanship held constant. Past work has shown that the difference between the total effect and the controlled direct effect can be interpreted as a combination of an indirect (or mediated) effect and a causal interaction (VanderWeele, 2015). As we argue in this paper, both are components of a causal mechanism, meaning that, when this difference is large, we can infer that the mediating attribute helps explain the overall effect, and

thus plays a role in the mechanism of that treatment. In describing the design, we compare this approach to a traditional mediation analysis that focuses only on indirect effects. We also highlight the trade-off inherent in our approach: while our assumptions are weaker than those needed for mediation, they cannot separately identify the indirect effect and the causal interaction. Nevertheless, our proposed quantities of interest still provide valuable information about causal mechanisms, more broadly defined (see also Gerber and Green, 2012, Ch. 10).

Our second contribution is to show how this approach to direct effects and causal mechanisms is affected by imperfect manipulation of the mediator. In survey experiments, the key mediating variable is often not necessarily the *provision* of some key piece of information, but rather the respondent's *belief* about that information. If these differ, then the average controlled direct effect of the mediating variable (the belief) will not be identified and the standard decompositions that we discussed above will not apply. To address this, we introduce a novel set of sufficient assumptions to recover a decomposition in this setting. We also show how to interpret results under imperfect manipulation of the mediator. Under our assumptions, we can use the manipulation (i.e., what the researcher tells the respondent) rather the mediating variable (i.e., what the respondent believes) and still recover a combination of the indirect and interaction effects of the mediating variable itself, with a slight change to interpretation of the interaction effect.

Our third contribution is to provide guidance on how intervening on a potential mediator can be (and, indeed, has been) applied in experimental settings, particularly in survey experiments. We demonstrate this using two illustrative examples. The first examines how the public evaluates nominees to the U.S. Supreme Court and documents how showing the respondents information about the nominee's partisanship reduces the signal conveyed to the respondents by the nominee's race or ethnicity (a topic explored in Sen, 2017). That is, most of the total effect of race can be explained by the inferred partisanship of the nominee. The second example replicates findings from Tomz and Weeks (2013) on the theory of the "democratic peace" 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 the potential threat of a nuclear program plays a role in how Americans decide to support preemptive strikes against democracies versus non-democracies. Importantly, we reach this conclusion without requiring the strong assumptions of the original paper's mediation analysis.

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, and describe how our approach differs from existing frameworks. Next, we define our two main causal quantities of interest: (1) controlled direct effects and (2) natural mediator effects. We show how these quantities can be defined and identified under both perfect and imperfect manipulation of the mediator. Furthermore, we explain how these quantities apply not just to experiments (and survey experiments in particular, using our illustrative example), but also more broadly to observational contexts. We then analyze the two above experiments, both of which show that we can identify different quantities of interest depending on the information provided to respondents. We conclude with a discussion of the implications of our study for applied researchers using survey experiments.

## 2 Setting and Illustrative Example

We develop the main ideas using the example of a candidate choice survey experiment. Suppose a researcher is interested in understanding how the public evaluates potential U.S. Supreme Court nominees; specifically, the researcher is interested in understanding to what extent 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 simply with a candidate identified to the respondents as African American and another simply with a candidate identified as white.

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 racial animus. Or, a negative treatment effect among certain respondents could also be attributed to a prior belief that black nominees are more likely to be Democrats or liberal (McDermott, 1998). Yet another possibility is that such a 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 different substantive conclusions: the first mechanism relies on racial prejudice while the second and third use race as a heuristic for other characteristics.

Manipulating the information environment can help researchers investigate these differing explanations. To illustrate, if the researcher included information about the candidate's partisanship in his experiment (as part of the candidate's profile, for example) then he would be able to assess whether the second hypothesis has support. If he included information about the candidate's professional background in the survey experiment, then he would be able to assess support for the third hypothesis. This kind of approach—increasingly popular in 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, researchers using these kinds of designs want to estimate the baseline causal effects of each attribute. Looking at our example again, does an effect of nominee race on respondent choice exist? This kind of question 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. More complicated is the second goal, which is that, given a particular total effect (or marginal component effect, to use 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? This type of question has been of increasing interest across social science, but most researchers look-

ing at these questions have proceeded in an *ad hoc* basis. Our goal here is to reason more formally about this second goal—that of investigating mechanisms.

## 2.1 Mechanisms, Mediation, and Interaction

We turn now to explaining what we mean by a causal mechanism and how certain experimental designs facilitate their exploration. 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 how an intervention or a change in contextual forces could have produced a different result. Building from the framework introduced by VanderWeele (2015), we define a *causal mechanism* as either a description of (1) the causal process, or how a treatment affects an outcome, or (2) a causal interaction, or in what context does the treatment affect the outcome. We note that past approaches to causal mechanisms, such as Imai et al. (2011), have equated causal mechanisms with indirect effects and causal processes, exclusively. But we believe that 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 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 (Imai et al., 2011). 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 illustration of black versus white Supreme Court nominees, this could be how the *race* of the hypothetical nominee 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 (VanderWeele, 2015, p. 9). In terms of 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 (or less) salient consideration when the nominee is of the same party as the respondent.

We note 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 Supreme Court nominee profiles with different racial backgrounds and also with different partisan affiliations (i.e., with randomly assigned combinations of the two). By design, race (the treatment) does not causally affect partisanship (the mediator) because both have been randomly assigned. However, the effects of race on respondent evaluation of the hypothetical nominee may still nonetheless depend on the value taken by partisanship (the mediator). Moreover, the interactions between the two, as we discussed above, yield insights into the mechanism by which race affects respondents' evaluations in situations where partisanship is not manipulated. We still use the language of "mediator" since these factors may mediate the effect when not manipulated. Below we also consider the case where the researcher can only imperfectly manipulate the mediator.

**Differences with other approaches.** Our approach differs in some important respects from existing frameworks. For example, 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 a 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 respondents. 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. We do engage in the beliefs of respondents below imperfect manipulation of the mediators.

Another approach is that of Imai, Tingley and Yamamoto (2013), who explore various experimental designs (including the one we consider below) that help identify mediation effects and thus focus on mechanisms as causal processes. 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 unpacking a causal mechanism in terms of processes and interactions is impossible. It also motivates our present set of questions—what can we learn or explain about a set of causal effects from these experimental designs?

Perhaps most similar to our approach is that of Gerber and Green (2012), who propose an "implicit mediation analysis," which involves creating multiple versions of the treatment that differ in theoretically meaningful ways and can provide insight into causal mechanisms (pp. 333–6). The approach we take in this paper is a version of this implicit mediation analysis, but we extend their ideas to discuss exactly what quantities of interest can be identified and how those might speak to specific causal questions. Below, we also build on the analysis of "manipulating the mediator" experiments in Gerber and Green (2012), addressing their concerns about the inability of a researcher to set values of the mediator perfectly.

[1] For example, using our illustration, if the researcher only provided respondents with information about the candidate's race (and not about partisanship), 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.

# 3 Assumptions and Quantities of Interest

We now present the formalism. We denote the treatment by $T_i$, where $T_i$ can take on one of $J_t$ values in the set $\mathcal{T}$. To keep the discussion focused, we assume that there is only one attribute in $T_i$ (such as race in our example), but below we discuss extending the framework to handle a multidimensional treatment, as in a conjoint design. There is also a potential mediator, $M_i$, which we assume is binary. (We address multi-leveled mediators in the Supplemental Materials.) In our example, $T_i = 0$ would indicate that a hypothetical Supreme Court nominee was reported to be African American and $T_i = 1$ would indicate that the nominee was reported to be white. The mediator might be partisanship; for example, $M_i = 0$ would indicate that the nominee is a Democrat and $M_i = 1$ that the nominee is a Republican.

We consider a setting with parallel survey experiments, which we indicate by $D_i \in \{d_*, d_0, , d_1\}$, where $i$ is the subject (Imai, Tingley and Yamamoto, 2013). Subjects with $D_i = d_*$ are in the *natural-mediator arm*, in which only the treatment is randomized. In the other arms, called *manipulated-mediator arms*, both the treatment and the mediator are randomized for subject $i$. For example, $D_i = d_1$ represents informing the subject that the nominee is a Republican (and so $M_i$ should be 1) and $D_i = d_0$ represents informing the subject that the nominee is a Democrat (and so $M_i = 0$).

To define the key quantities of interest, we rely on the potential outcomes framework for causal inference (Rubin, 1974; Holland, 1986; Neyman, 1923). In this setting, the mediator has potential outcomes that possibly depend on both the treatment and experimental arm, $M_i(t, d)$, which is the value that the mediator would take for subject $i$ if they were assigned to treatment condition $t$ and experimental arm $d$. For example, $M_i(t, d_*)$ would be the value that the mediator (partisanship) would take if the respondents were given information only about nominee race.[2] In the manipulated-mediator arm with $D_i = d_0$, on the other hand, both the treatment and the mediator would be assigned by the researcher. This would correspond in our example with providing respondents with race/ethnic information *and* partisan information about the hypothetical nominees. For now, we

[2] In this case, respondents may assume that a nominee identified as black is a Democrat (McDermott, 1998). Such a presumption would be in line with what Dafoe, Zhang and Caughey refer to as confounding.

assume *perfect manipulation of the mediator* so that $M_i(t, d_1) = 1$ and $M_i(t, d_0) = 0$ for all respondents and all levels of treatment, $t$. That is, we assume that if we tell the subjects that the nominee is a Democrat, $D_i = d_0$, then the subject believes the candidate is a Democrat, $M_i = 0$. Below, we weaken this assumption to allow for imperfect manipulation of the mediator.

In each experiment, the subjects have potential outcomes associated with every combination of the treatment and the mediator, $Y_i(t, m, d)$, 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 the potential outcome evaluated at the observed combination of the treatment and the mediator. As in Imai, Tingley and Yamamoto (2013), we make the following exclusion restriction:

**Assumption 1** (Manipulation Exclusion Restriction). *For all* $(t, m) \in \mathcal{T} \times \mathcal{M}$ *and* $(d, d') \in \{d_*, d_0, d_1\}^2$,

$$Y_i(t, m, d) = Y_i(t, m, d') \equiv Y_i(t, m).$$

The assumption states that the experimental arm only affects the outcome through its influence on the value of the mediator. In our example, this means that we assume a respondent's support for the candidate is the same regardless of whether the respondent infers that the nominee is a Democrat from the racial information as opposed to whether she was actually provided with the explicit cue that the nominee is a Democrat. This assumption could be violated if, for example, giving the respondents partisan information leads them to presume the study itself is about partisanship, thereby causing them to put increased importance on partisanship in that context and not in the other experimental arms where it is not provided.

The exclusion restriction 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 $Y_i(t) = Y_i(t, M_i(t, d_*))$ to be the potential outcome just setting the value of the treatment. We also make a consistency assumption that connects the observed outcomes to the potential outcomes, such that $Y_i = Y_i(T_i, M_i)$ and $M_i = M_i(T_i, D_i)$.

We make a randomization assumptions that follows directly from how these experiments are usually designed. We assume that both the treatment and the experimental-arm indicator are randomly assigned:

**Assumption 2** (Parallel Randomization). *For all* $(t, t', m, d) \in \mathcal{T}^2 \times \{0, 1\} \times \{d_*, d_0, d_1\}$,

$$\{Y_i(t, m), M_i(t', d)\} \perp\!\!\!\perp \{T_i, D_i\}$$

This assumption implies that the treatment alone is randomized in the in the natural-mediator arm and that both the treatment and the mediator are randomized in the manipulated-mediator arm. In extending this analysis to observational data, these assumptions can be generalized to accommodate covariates, both pretreatment and intermediate (see, for instance, Acharya, Blackwell and Sen, 2016).

## 3.1 Quantities of Interest: Indirect, Interaction, and Natural Mediator Effects

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

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

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 arms allow us to analyze the joint effect of intervening on both the treatment and the mediator. 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 example involving Supreme Court nominees, the total treatment effect is the difference in support for a hypothetical black candidate versus a white candidate for unit *i*. The controlled direct effect, on the other hand, would be the difference in support between these two nominees where respondents are provided with the additional information that the two nominees are of the same 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 $CDE(t_a, t_b, m) = \mathbb{E}[CDE_i(t_a, t_b, m)] = \mathbb{E}[Y_i(t_a, m) - Y_i(t_b, m)]$. As we discuss below, the controlled direct effect can be thought of as the part of the total effect that is due to neither mediation nor interaction with $M_i$ (VanderWeele, 2014).

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

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

This is the effect of changing the mediator with a change in treatment, but keeping treatment fixed at a particular quantity. (In our example, this could be the difference in respondent's support when the candidate is black versus support when the candidate is black, but the partisanship is set to level the respondent would infer *if the candidate were white*.) In practice, the second term in the effect, $Y_i(t_a, M_i(t_b, d_*))$, 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, identification of this quantity will often require strong and perhaps unrealistic assumptions. 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 $M_i(t_a, d_*) = M_i(t_b, d_*)$, or (2) the mediator has no effect on the outcome. It is intuitive that the NIE would be equal to zero under either condition, given the usual motivation of indirect effects as multiplicative: the effect of treatment on the mediator is multiplied by the effect of the mediator on the outcome.[3] As above, we define the *average natural indirect effect* (ANIE) to be $NIE(t_a, t_b) = \mathbb{E}[NIE_i(t_a, t_b)]$.

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

**Reference Interactions.** To capture the interaction between $T_i$ and $M_i$, we introduce the so-called *reference interaction* (VanderWeele, 2014), which is the difference in controlled direct effects between the reference category $m$ and the natural value of the mediator under $t_b$, or $M_i(t_b, d_*)$:

$$RI_i(t_a, t_b, m) = \mathbb{I}\{M_i(t_b, d_*) = 1 - m\} \left[CDE_i(t_a, t_b, 1 - m) - CDE_i(t_a, t_b, m)\right] \tag{4}$$

In our example, the reference interaction would compare the CDEs of black versus white nominees at two levels: (1) the inferred partisanship under a white nominee and (2) the manipulated partisanship (for example, party set to "Republican"). 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:

$$\begin{aligned} &RI(t_a, t_b, m) = \mathbb{E}[RI_i(t_a, t_b, m)] \\ &= \mathbb{E}\big[CDE_i(t_a, t_b, 1 - m) - CDE_i(t_a, t_b, m) | M_i(t_b, d_*) = 1 - m\big] \mathbb{P}[M_i(t_b, d_*) = 1 - m] \end{aligned} \tag{5}$$

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). When $m = 0$ (in our case, when the candidate is revealed to be a Democrat), then this quantity is the average change in the CDE between Republican and Democratic candidate profiles for those who naturally think the candidate is a Republican, weighted by the size of this latter group. The ARIE provides a summary measure of how the ACDE varies across units due to variation in the mediator—it is the part of the total effect that is due to interaction alone (VanderWeele, 2014). 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 natural 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. This quantity may be equal to zero if there are exact cancellations in the interactions across the population, but this is both rare and dependent on the baseline category, $m$. In the Supplemental Materials, we show that when the mediator is multileveled, the reference interaction has a similar form: it is the weighted average of interactions across all levels of the mediator (compared to $m$), weighted by the probability that the natural mediator takes those values.

One drawback of the ARIE is that it is dependent on the baseline or reference category $m$. That is, the ARIE for setting the partisan label of the nominee to "Democrat" will differ from the ARIE setting

it to "Republican." In fact, the sign of these two effects may be different, making careful interpretation of this quantity essential. As a practical matter, it is often useful to set $m = 0$, so that the interpretation of the ARIE is with regard to *positive* changes in $M_i$. These concerns are very similar to issues of interpreting interactions in many statistical models, including linear regression.

**Natural Mediator Effects.** The proposed design involves the choice to intervene on the mediator or not, lending us to introduce another quantity of interest, the *natural mediator effect*, or NME. The natural mediator effect is the effect of changing the mediator to its natural value for a particular treatment value relative to some fixed baseline level of the mediator:

$$NME_i(t, m) = Y_i(t) - Y_i(t, m) = Y_i(t, M_i(t, d_*)) - Y_i(t, m) \tag{6}$$

Another way to understand this quantity is as the negative of an intervention effect—that is, the effect of intervening on the mediator and setting it to some value. This quantity is o if the natural level of the mediator under $t$ is equal to the baseline value, so that $M_i(t, d_*) = m$ or if the mediator has no effect on the outcome. Intuitively, the NME is the effect of the induced or natural level of the mediator under treatment level $t$ relative to $m$. This quantity is often of interest for applied researchers. To provide intuition, consider a study looking at the effects on weight gain of two prescriptions: diet and exercise. Natural mediator effects would be appropriate if a researcher was interested in how weight changes when subjects with the same assigned level of exercise are allowed to choose their own diet (which would likely cause people to eat more) relative to a fixed prescription of both diet and exercise. Specifically, in this case, the researcher would be interested in knowing the effect of the natural level of the diet under a particular exercise regime. Using our illustration of Supreme Court nominees, the natural mediator effect would be the effect of inferred (natural) partisanship of a hypothetical black nominee relative to a baseline value of that candidate being a Democrat. Some respondents will infer the partisanship of the hypothetical candidate to be a Democrat, which implies that the NME will be zero for those respondents since, for them, $M_i(t, d_*) = m$. Unlike the indirect effect, though, the NME might be non-zero even if there is no effect of treatment on the mediator, since $M_i(t_a, d_*) = M_i(t_b, d_*) \neq m$. This would be true for respondents who always infer the candi-

date to be a Republican, no matter their stated race. The *average natural mediator effect* (ANME) is $NME(t, m) = \mathbb{E}[NME_i(t, m)] = \mathbb{E}[Y_i(t) - Y_i(t, m)]$, and it is a suitable quantity of interest for experiments that provides additional information to some, but not all respondents. This may be the case in conjoint experiments, vignette experiments, or certain field experiments where the intervention involves manipulating the information environment.

**Difference in Natural Mediator Effects.** The NME gives us some intuition about how subjects respond to the mediator when we move from a controlled mediator to its natural value under a particular treatment. But the notion of a causal mechanism of a treatment is necessarily about comparisons across treatment levels. Thus, the *difference in natural mediator effects* (DNME) is

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

This quantity tells us how the natural mediator effect varies by level of the treatment. It will be equal to o whenever there is no effect of the mediator on the outcome at any level of treatment and thus no causal interaction or indirect effect. The DNME is equivalent to the difference between the treatment effects in the natural-mediator and manipulated-mediator arms:

$$\Delta_i(t_a, t_b, m) = \underbrace{\left[Y_i(t_a, M_i(t_a, d_*)) - Y_i(t_b, M_i(t_b, d_*))\right]}_{\text{total effect}} - \underbrace{\left[Y_i(t_a, m) - Y_i(t_b, m)\right]}_{\text{controlled direct effect}} \tag{8}$$

That is, this quantity is also the difference between the total treatment effect and the controlled direct effect at mediator value $m$. In the context of the Supreme Court nominee experiment, this would tell us the effect of inferred partisanship versus manipulated partisanship (e.g., party set to Democrat) for black nominees compared to the same effect for white nominees—a type of difference-in-differences quantity. 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 nominees are Democrats. Finally, because each of the NMEs can be seen as (the negative of) an intervention effect as described above, we can also think of this quantity as a difference in intervention effects—that is, it would be the difference in effect of intervening on the mediator at two levels of treatment. The *average*

*difference in natural mediator effects* (ADNME) is $\Delta(t_a, t_b, m) = \mathbb{E}[NME_i(t_a, m) - NME_i(t_b, m)] = NME(t_a, m) - NME(t_b, m)$, which is simply the difference in average natural mediator effects at two levels of the treatment given the manipulated level of the mediator $m$.[4]

## 3.2 How Natural Mediator Effects Help Us Understand Causal Mechanisms

In this section, we explain how the difference between natural mediator effects can teach us about the underlying causal mechanisms. Under consistency, we can characterize the difference in natural mediator effects (or the difference between the total and controlled direct effects) using the following decomposition (VanderWeele, 2014; VanderWeele and Tchetgen Tchetgen, 2014):

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

The difference between the total effect and the controlled direct effect, 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 is thus a combination of the two aspects of a causal mechanism: (1) the causal process, represented by the indirect effect, and (2) the causal interaction, represented by the interaction effect. Thus, we can interpret the ADNME as the portion of the ATE that can be explained by $M_i$, either through indirect effects or interactions.

In the Supreme Court nominee example, the difference in intervention effects is the combination of two separate components. The first is the indirect effect of race on choice through partisanship. The second is the interaction between partisanship and race, for those units that would think a white nominee ($t_b$) is a Republican, $M_i(t_b, d_*) = 1$, 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, this latter condition may be plausible. For example, given that few African Americans identify as Republicans, assuming

[4] Under a different set of assumptions, Robins and Greenland (1992) referred to this quantity as the "effect that could be eliminated by controlling for" $M_i$ (p. 152). When divided by the average treatment effect, VanderWeele (2015, p. 50) calls this the "proportion eliminated." Both of these names reflect the original use of these quantities under certain monotonicity assumptions. We find this naming can be confusing when, for example, the ACDE is greater in magnitude than the ATE and so the "effect eliminated" can be greater than the original overall effect. For these reasons, we opt for the ADNME naming convention.

that nearly all respondents would infer such a nominee to be a Democrat may be reasonable. In these cases, the difference in the intervention effects can be interpreted as, essentially, the indirect effect. We note that even when these conditions do not hold, the ADNME 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, disentangling the relative contribution of the indirect and interaction effects in contributing to the difference in natural mediator effects is impossible. In order to do so, we require stronger assumptions such as a no interaction between $T_i$ and $M_i$ at the individual level or independence between the natural value of the mediator and 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 $CDE_i(t_a, t_b, m_c) - CDE_i(t_a, t_b, m_d) = 0$ which implies that the reference interaction must be 0 and the difference in natural mediator effects is exactly equal to the indirect effect (Robins, 2003). This approach is problematic because such "no interaction" assumptions are highly unrealistic in most settings (Petersen, Sinisi and van der Laan, 2006). Imai, Keele and Yamamoto (2010) show how independence between the natural value of the mediator and the outcome allows one to identify the indirect effect separately from the interaction, but this independence is a strong assumption that can be violated in empirical examples. The approach in this paper makes weaker assumptions, but can only identify a combination of the indirect and interaction effects. Thus, there exists a fundamental trade-off between the strength of the assumptions maintained and ability to distinguish between indirect effects and interactions. Fortunately, all is not lost when the mediation assumptions fail to hold: with a broader view of causal mechanisms, such as the one we suggest here, the ACDE and the proposed design can still provide useful, albeit coarse, evidence about mechanisms.

## 3.3 Imperfect manipulation of the mediator

Up to this point, we have assumed the mediator of interest could be manipulated, which is a reasonable assumption in survey experiments where the mediator is the actual *provision* of information. But if researchers want to treat the *belief* of this information as the mediator, then the above analysis is incomplete. In our example, respondents might not believe a nominee is a Democrat when informed

in the experiment the nominee is Democrat—particularly if different respondents form different beliefs based on the same information. In the example of diet and exercise, participants assigned to a specific combination of diet and exercise might cheat on their diet, eating more than the assigned amount. The goal in this section is to outline the assumptions necessary to learn about the causal mechanisms associated with the "true" mediator even when we cannot directly affect it.

We introduce the following assumptions that put structure on the manipulations:

**Assumption 3** (Imperfect Manipulation). *For all* $t \in \mathcal{T}$*:*

1. *Monotonicity:* $M_i(t, d_0) \leq M_i(t, d_*) \leq M_i(t, d_1)$*;*
2. *Manipulation crowd-out:* $M_i(t, d) = M_i(t', d) = M_i(d)$ *when* $d \in \{d_0, d_1\}$.

Monotonicity states that providing information does not have perverse effects. For example, suppose that $d_0$ here refers to "Democrat" and $d_1$ corresponds to "Republican," where treatment is still the race of the candidate. Monotonicity rules out pathological cases where under no manipulation the respondent believes a candidate is a Democrat ($M_i(t, d_*) = 0$), but when told that the candidate is a Democrat would believe that the candidate is a Republican ($M_i(t, d_0) = 1$). The second part of the assumption is that when information is provided about the mediator, it is sufficiently strong that it crowds out any effect of treatment. In other words, this assumes that the treatment is unrelated to the noncompliance with the mediator manipulation. Robins and Greenland (1992, p. 149) considered stronger versions of these assumptions to identify indirect effects, but their approach maintained a no-interactions assumption.

When we cannot directly manipulate the mediator, we can no longer identify the ACDE with $M_i$ fixed as some value. To address this, we define an alternative version of the ACDE with the experimental arm fixed, $D_i = d_0$, instead of the mediator:

$$CDE^*(t_a, t_b, d_0) = \mathbb{E}[Y_i(t_a, M_i(d_0)) - Y_i(t_b, M_i(d_0))] \tag{10}$$

This is the estimand that would be identified in the manipulated mediator arm under imperfect manipulation, so long as the exclusion restriction and randomization hold. We can also define similarly

modified versions of the natural mediator effects, $NME^*(t, d_0) = \mathbb{E}[Y_i(t) - Y_i(t, M_i(d_0))]$, and the difference in natural mediator effects, $\Delta^*(t_a, t_b, d_0) = NME^*(t_a, d_0) - NME^*(t_b, d_0)$. These effects are now defined in terms of the experimental arm manipulation rather than the mediator directly. To see how the decomposition results above change in this setting, let $L_i = 1$ if respondent $i$ can have their minds changed when provided information $d_0$. This group would believe that the mediator is at the high value without the manipulation, $M_i(t_b, d_*) = 1$, but would change their mind if given information $d_0$, $M_i(d_0) = 0$. Then, we show in the Appendix that the following decomposition holds:

$$\Delta^*(t_a, t_b, d_0) = TE(t_a, t_b) - CDE^*(t_a, t_b, d_0) \tag{11}$$

$$= NIE(t_a, t_b) + \mathbb{E}[CDE_i(t_a, t_b, 1) - CDE_i(t_a, t_b, 0)|L_i = 1]\mathbb{P}[L_i = 1] \tag{12}$$

The difference between the total effect and the controlled direct effect (at level $d_0$) is the sum of the indirect effect and a modified interaction effect. This modified interaction effect is the difference in the CDE between the low and high value of the mediator *for those who update their beliefs in response to manipulation*, weighted by the the size of this group. This is similar interpretation to reference interaction under perfect manipulation and only differs in which groups are omitted from the interaction effect. Under imperfect manipulation, for instance, there are respondents who always believe the candidate is Republican no matter what they are told, $M_i(d_1) = M_i(t, d_*) = M_i(d_0) = 1$, but they are omitted from the interaction effect. Perfect manipulation of the mediator rules this group out entirely, so they are obviously absent from the interaction effect in that setting as well. Thus, we can conclude that the monotonicity and crowd-out assumptions above are sufficient for interpreting the difference in natural mediator effects as if there was perfect manipulation of the mediator. In the Appendix, we derive a decomposition without the crowd-out assumption, in which case there is a change to the interpretation of the indirect effect as well.

## 3.4 Extension to Conjoint Experiments

The above framework can be easily extended to conjoint experiments where several attributes are manipulated at once and several separate profiles are shown to each respondent, as is done in con-

joint experiments. This would mean that $T_i$ is actually a multidimensional vector indicating the set of profiles provided to respondent $i$. For example, our treatment might include information about the race of the proposed Supreme Court 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 assumptions of no-profile order effects and no carry-over effects, simple difference-in-means estimators that aggregate across respondents are 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 conjoint experiments, we can replace the ATE in the above discussion with the AMCE and much of interpretation remains intact. This allows us to think of the difference in natural mediator 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 (or lack thereof) in the profile helps explain the "total" effect of the AMCE.

## 3.5 Relationship to Post-Treatment Bias

When thinking about variables possibly affected by the main treatment of interest, a common threat to inference is *post-treatment bias* (Rosenbaum, 1984). Post-treatment bias can occur when one conditions on a variable that is affected by the treatment (making it "post-treatment"). It is useful to partition this bias into two different types that are often conflated. First, conditioning on a post-treatment variable will generally change the quantity of interest under study from the ATE to the ACDE, which is often the goal of such an approach. Second, conditioning on a post-treatment variable can induce selection bias (sometimes called "collider bias") that will bias most estimators away from either the ACDE or the ATE. Luckily, in the framework presented here, neither of these cause any problems. The first type of post-treatment bias is actually our target of estimation here—the difference between the ATE and ACDE. And, because the studies we consider here experimentally manipulate the mediator, selection bias does not arise here. In observational studies, on the other hand, post-treatment bias

can arise when attempting to control for factors that confound the mediator-outcome relationship (Acharya, Blackwell and Sen, 2016).

### 3.6 Relevance for Observational Studies

Our approach also relates to observational studies and to the approach taken by Acharya, Blackwell and Sen (2016). Thinking of observational studies as having experimental interpretations illustrates the logic: for example, what is the hypothetical experiment that would identify the causal parameter of interest in the observational study? In cases where the average treatment effect and the controlled direct effect are both identified in an observational study, the decomposition in (8) implies that we can also identify the ADNME. Acharya, Blackwell and Sen (2016) proposed the difference between the ATE and the ACDE as a measure of the strength of a mechanism; this difference has a straightforward interpretation as the difference in natural mediator effects from the above experimental design. The estimation and inference for those observational studies is often more complicated than the above experimental setting because of the presence of both baseline and intermediate confounders.

The fact that the treatment effect can be decomposed into a controlled direct effect and difference in natural mediator effect suggests that the ADNME 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 natural mediator effects in this example is the following. First consider intervening on a unit where income is set to a pre-specified level and then varying the level of plough use from that pre-specified level to the natural, realized 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 of plough use between these two cases. If the two natural mediator effects are the same, we

might interpret this as evidence that contemporary income does not "explain" the effect of historical plough use, either through mediation or interaction. However, if they are different, we might interpret it as evidence that it does explain some (or all) of it. While these interventions are obviously hypothetical, they highlight the relevant counterfactuals in observational studies like this one.

# 4 Estimation

We now turn to identification and estimation strategies. Under the assumptions above, we can show that the difference in natural mediator effects under imperfect manipulation of the mediator is identified as:

$$\begin{aligned}\Delta^*(t_a, t_b, d_m) = &\big[\mathbb{E}[Y|T_i = t_a, D_i = d_*] - \mathbb{E}[Y|T_i = t_a, D_i = d_m]\big] \\ &- \big[\mathbb{E}[Y|T_i = t_b, D_i = d_*] - \mathbb{E}[Y|T_i = t_b, D_i = d_m]\big]\end{aligned} \tag{13}$$

We omit a proof given that it would be a straightforward application of standard results in experimental design. Note that under perfect manipulation of the mediator, we have $\Delta(t_a, t_b, m) = \Delta^*(t_a, t_b, d_m)$, so this expression also identifies the difference in natural mediator effects that in that setting as well.

How might we estimate this quantity with our experimental samples? A simple plug-in estimator would replace the expectations above with their sample counterparts. For instance, we would estimate $\mathbb{E}[Y_i|T_i = t_a, D_i = d_*]$ with:

$$\widehat{\mathbb{E}}[Y_i|T_i = t_a, D_i = d_*] = \frac{\sum_{i=1}^{N} Y_i \mathbb{I}\{T_i = t_a, D_i = d_*\}}{\sum_{i=1}^{N} \mathbb{I}\{T_i = t_a, D_i = d_*\}} \tag{14}$$

Replacing each of the expectations in (13) in a similar fashion would produce an unbiased estimator for $\Delta$. A convenient way to produce this estimator is through linear regression on a subset of the data. Specifically, to estimate these quantities, first let $Z_i$ be an indicator for the natural mediator arm—that is, $Z_i = 1$ when $D_i = d_*$. It is sufficient to subset the manipulated-mediator arm with mediator value $m$ ($D_i = d_m$) 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, $Z_i$, and interactions $W_{it}Z_i$. Under perfect manipulation of the mediator, if $t_b$ is the omitted category, then the coefficient on $W_{it_a}$ is an unbiased estimator of $CDE(t_a, t_b, m)$ and the coefficient on $W_{it_a}Z_i$ will be equivalent to the above nonparametric estimator

for the ADNME, $\Delta(t_a, t_b, m)$. 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, $D_i = d_*$.

Estimation with conjoint experiments under complete randomization across and within experimental arms is straightforward. Let $T_{ikl}$ represents the $l$th attribute of the $k$th profile being evaluated, which can take on $J_l$ possible values, and let $Y_{ik}$ is subject $i$'s response to the $k$th 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 natural mediator effects for a particular attribute, we simply interact these dummy variables with the experimental-arm indicator, $Z_i$. 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. For more details on estimation in conjoint experiments, see Hainmueller, Hopkins and Yamamoto (2013).

# 5 Experimental Analysis of Direct Effects and Mechanisms

## 5.1 Study #1: Conjoint Experiment for Nominees to the U.S. Supreme Court

As a first application of these ideas, we take an example from Sen (2017) on how the public views nominees to the U.S. Supreme Court. The experiment provides an attractive illustration for the reason that the true ideological leanings of Supreme Court nominees is often noisily conveyed to the public. For that reason, the data are unique in the sense that half of the 1,650 respondents in the study were randomly assigned to see a conjoint profiles that contained partisan information about a potential nominee ($n = 886$) and half were assigned to see profiles that contained no such information ($n = 764$). The outcome variable is a binary measure of support of the nominee.

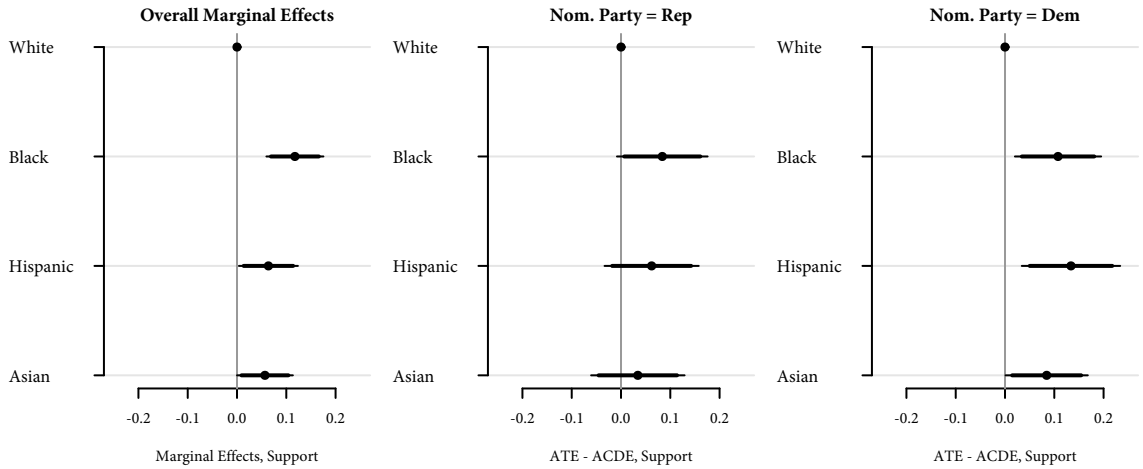


Figure 1: Average marginal effects of nominee race on support for the nominee (left panel) and average difference in natural mediator effects for nominee partisanship as a mediating variable (middle and right panels). The middle panel has the partisanship in the manipulated-mediator arm set to "Republican" and the right panel has the partisanship set to "Democrat." All effects are relative to the baseline of a white nominee. Thick and thin lines are 90% and 95% confidence intervals, respectively.

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 Democratic or Democratic 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 another 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 natural mediator 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 could estimate the AMCE and the ADNME for each attribute of the conjoint profiles.

Thus, in the natural-mediator arm of the experiment, respondents rated profiles that included race, gender, age, religion, previous work experience, and law school rank, but excluded any infor-

mation about the nominee's partisanship. In the manipulated-mediator arm, the profiles included information about the party affiliation of the nominee in addition to all of the attributes. We focus on the 583 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 manipulated-mediator arm of the experiment. To analyze this experiment, we estimate the AMCE from the natural mediator arm, then estimate the ACDE from the manipulated-mediator arm, and then use these two quantities to estimate the ADNME. (In other words, this is the difference in the effect of going from black to white under no party information versus party being set to Republican *or* Democrat.[5]) Given the above discussion, these ADNMEs will give us some sense of whether partisanship participates in a mechanism for the various marginal effects of each component.

Figure 5.1 shows the results for the effects of nominee race, with the total AMCEs in the left panel and the ADNMEs for Republican profiles and Democratic profiles in the middle and right panels, respectively. The Figure shows both 95% and 90% 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. (In Appendix Figure A1, we show the full set of component effects, which show that these respondents are also more likely to support nominees that served as a law 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 the effect of race on support due to respondents inferring the partisanship of the nominee from their race? The differences in natural mediator effects tell us this exactly. In the right two panels, we show the ADNMEs when setting the candidate party to two different levels, Republican and Democrat. A large, statistically significant difference for a given attribute in either of these panels would indicate that partisanship of the hypothetical candidate plays a role in a causal mechanism for that attribute. The ADNMEs for 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

[5] Note that there are two possible ACDEs, one for Republican (non-copartisan) and Democratic (copartisan) profiles and so there are two possible ADNMEs corresponding to each of these.

especially acute for the effect of a black nominee versus a white nominee, which makes sense since black citizens are more likely to identify with the Democratic party than white citizens. In the Appendix, we show that partisanship plays less of a role for the other attributes with a few exceptions. The above interpretations continue to hold even if not everyone believes a candidate to be a member of the party provided, so long as the partisan affiliation manipulation crowds out the effect of race as described in Section 3.3.

These differences imply that there are either indirect effects of race on support through inferred partisanship or that there are positive interactions between racial attributes 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 see in the natural-mediation arm, which is consistent with the literature on heuristics from political psychology (e.g., McDermott, 1998). Finally, we note that this is a study where the possibility of conducting a mediation analysis might be fraught. The sequential ingorability assumption of Imai, Keele and Yamamoto (2010) would require us to measure the inferred party of the nominee and then assume that this inferred partisanship is essentially randomly assigned with respect to the potential levels of support. This may be implausible in this case, making our design an attractive alternative to learning about the causal mechanisms.

## 5.2 Study #2: Public Opinion and Democratic Peace

As a second application of this framework, 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 presented respondents with different country profiles and asked respondents whether they would, or would not, support preemptive American military strikes against the hypothetical country. 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 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 a Mechanical Turk sample of 1,247 respondents,[6] 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 manipulated 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."[7] 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 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), but with some caveats. For example, the difference is not statistically significant, which might be due to the fact that the number of units used to estimate the ATE here is roughly half the number used in the original experiment. Also, the ACDE of democracy with the information about threat held constant 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

[6] 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.

[7] 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.

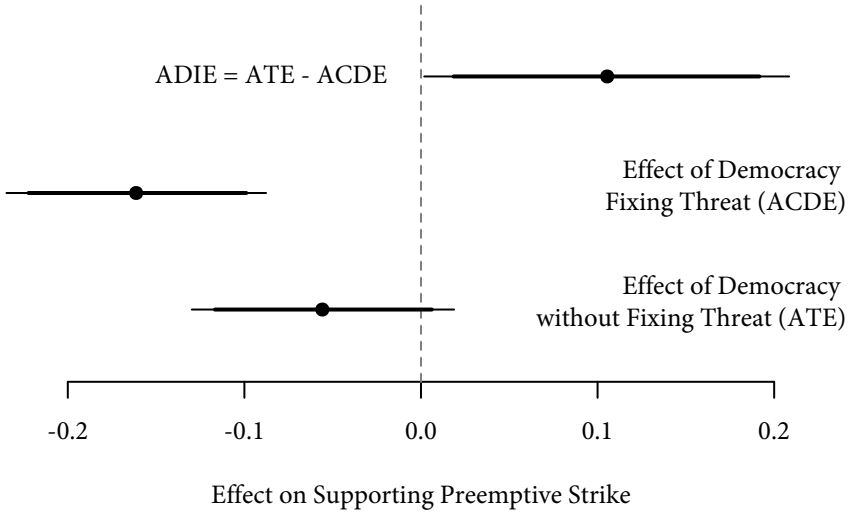


Figure 2: Results from the replication of Tomz and Weeks (2013). Data from a Mechanical Turk survey experiment ($N = 1247$). Bootstrap 95% (thin line) and 90% (thick line) confidence intervals are based on 5,000 bootstrap replications.

indirect effect of democracy through potential threat, which would imply that the natural direct effect should be smaller in magnitude (that is, less negative) than the ATE. Here we find the opposite—with a potential threat revealed, democracy has an even stronger negative effect on support for a strike.

There are two ways to reconcile these findings. First, recall that the ADNME (the difference between the ATE and the ACDE) is a combination of the indirect effect and the reference interaction effect. In our setup, the reference category for the ARIE could be seen as a relatively low level of threat since neither the U.S. nor its allies are being directly threatened. Thus, the ARIE in this case would represent the average change in the ACDE when moving to a higher level of threat, which we would expect to be positive if higher levels of threat caused the direct effect of democracy to attenuate. This interpretation is still consistent with threat being a mechanism for the total effect—the distribution of perceived threat levels under autocracy, $M_i(t_b, d_*)$, and its impact on the direct effect of democracy determine part of the overall ATE. This type of positive reference interaction could explain our results if it were larger in magnitude than any negative indirect effect, leaving the overall ADNME positive.

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 crucial sequential ignorability assumption

for mediation. 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. These two explanations are not mutually exclusive and could work together to produce the large, positive ADNME we see in this study. Either way, without further assumptions, it is impossible to tease apart the relative contributions of the indirect and interaction effects in this study. However, we can conclude that threat is part of a causal mechanism for the effect of democracy on support for a strike.

# 6 Conclusion

We conclude by providing an assessment of how our framework may be useful for applied researchers. Many of the most interesting political science questions focus 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, natural mediator effects, and differences in natural mediator 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" the possibility of an opposing narrative 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 his findings regarding the treatment effect of race. On the other hand, the natural mediator effect is perhaps a more intuitive step, as it represents the difference associated with intervening on a 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 natural mediator 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 best suits applied researchers' needs is the first step. The second is estimation. We provided a simple way to estimate the ACDE and the ADNME both in straightforward survey experiments and in more complicated conjoint designs. 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 natural mediator effects—making them particularly flexible for applied researchers.

Of course, the fairly weak assumptions of the proposed design come at a cost. Under the maintained assumptions, estimating the indirect effect of treatment separately from the interaction is impossible. Stronger assumptions, such as those proposed in Imai, Keele and Yamamoto (2010), allow for the identification of the indirect effect, which is an intuitive quantity of interest. Still, these additional mediation assumptions cannot be guaranteed to hold by experimental design and so could be false. Our goal in this paper is to highlight how we can still obtain evidence on causal mechanisms even when mediation assumptions are unlikely to hold. Applied researchers must evaluate what trade-offs are acceptable for each empirical 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.
**URL:** http://www.allandafoe.com/confounding

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. W.W. Norton.

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

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81(396):945–960.

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

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(04, November):765–789.

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1, February):51–71.

McDermott, Monika L. 1998. "Race and Gender Cues in Low-information Elections." *Political Research Quarterly* 51(4):895–918.

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.

Petersen, Maya L, Sandra E Sinisi and Mark J van der Laan. 2006. "Estimation of Direct Causal Effects." *Epidemiology* 17(3):276–284.

Robins, James M. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems*, ed. P. J. Green, N. L. Hjort and S. Richardson. Oxford University Press pp. 70–81.

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

Rosenbaum, Paul R. 1984. "The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society. Series A (General)* 147(5):656–666.

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

Sen, Maya. 2017. "How Political Signals Affect Public Support for Judicial Nominations: Evidence from a Conjoint Experiment." *Political Research Quarterly* . Forthcoming.
**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. 2014. "A Unification of Mediation and Interaction: A 4-Way Decomposition." *Epidemiology* 25(5, September):749–761.

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

VanderWeele, Tyler J and Eric J Tchetgen Tchetgen. 2014. "Attributing Effects to Interactions." *Epidemiology* 25(5, September):711–722.

# A Multilevel mediators

In this section, we generalize the discussion in the main text to allow for mediators with more than two levels. Now, the mediator can take values $m \in \mathcal{M}$, where there are $|\mathcal{M}| = J_m$ possible values $\{m_1, \ldots, m_{J_m}\}$. The potential values of the outcome and the mediator remain similarly defined as above and the assumptions should be slightly modified to hold for all values $m \in \mathcal{M}$. We also extend the notation of the manipulation variable so that $D_i$ takes on one of $J_m + 1$ values $\{d_*, d_1, \ldots, d_{J_m}\}$. We assume that the ordering of these values of $D_i$ are such that $d_k$ corresponds to setting $M_i = m_k$ and $d_*$ remains the natural-mediator arm. Finally, it is most intuitive to apply these methods to situations where there is an ordering to values of the mediator so that $m_{J_m}$ refers to a "higher" value of the mediator than $m_1$. With nominal mediators, it is often more useful to use a series of binary mediators.

All of the quantities of interest in the main text remain the same with this new mediator except for the reference interaction. For a given mediator level, $m$, the reference interaction at the individual level becomes:

$$RI_i(t_a, t_b, m) = \sum_{\tilde{m} \in \mathcal{M} \setminus \{m\}} \mathbb{I}\{M_i(t_b) = \tilde{m}\} \left[CDE_i(t_a, t_b, \tilde{m}) - CDE_i(t_a, t_b, m)\right] \tag{15}$$

Taking averages, we get the the ARIE in this setting:

$$\begin{aligned} &\mathbb{E}[RI_i(t_a, t_b, m)] \\ &= \sum_{\tilde{m} \in \mathcal{M} \setminus \{m\}} \mathbb{E}\big[CDE_i(t_a, t_b, \tilde{m}) - CDE_i(t_a, t_b, m) | M_i(t_b) = \tilde{m}\big] \mathbb{P}[M_i(t_b) = \tilde{m}] \end{aligned} \tag{16}$$

This new reference interaction can be interpreted as the weighted average interaction for those units whose natural value of the mediator is not equal to $m$. As with the binary case, this quantity 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 natural value of the mediator under $t_b$ being equal to anything

other than $m$. Under perfect manipulation of the mediator, the decomposition in (9) remains valid with this updated definition of the reference interaction (VanderWeele, 2015, p. 606).

# B Proof of imperfect manipulation decomposition

Here we develop the technical details of the imperfect manipulation with multileveled mediators and without the crowd-out assumption. The binary mediator results will follow. Below, we show how crowd-out simplifies our result. First, we state a more general version of the monotonicity assumption:

1. Monotonicity (i): $M_i(d_1) \leq \ldots \leq M_i(d_{J_m})$
2. Monotonicity (ii): $M_i(t, d_*) = m_j \implies M_i(d_j) = m_j$

Here, we extend monotonicity to rule out situations where a respondent naturally believes mediator to take value $m_j$, but changes that belief when they are told it is that value ($D_i = d_j$). In a binary setting, the second two assumptions are equivalent to Assumption 3. Finally, when crowd-out doesn't hold, there is a possibility of a indirect effect in the manipulated-mediator arms, so we define a more general natural indirect effect that can depend on the experimental arm: $NIE_i(t_a, t_b, d) = Y_i(t_a, M_i(t_a, d)) - Y_i(t_a, M_i(t_b, d))$.

We show the decomposition of the difference between the total effect and the controlled direct effect at the lowest level of the mediator, $d_1/m_1$:

$$Y_i(t_a) - Y_i(t_b) - [Y_i(t_a, M_i(t_a, d_1)) - Y_i(t_b, M_i(t_a, d_1))] \tag{17}$$

By adding and subtracting $Y_i(t_a, M_i(t_b, d_*))$ and $Y_i(t_a, M_i(t_b, d_1))$, we can see that this is equivalent to:

$$\begin{aligned} NIE_i(t_a, t_b, d_*) - NIE_i(t_a, t_b, d_1) \\ + [Y_i(t_a, M_i(t_b, d_*)) - Y_i(t_b, M_i(t_b, d_*))] \\ - [Y_i(t_a, M_i(t_b, d_1)) - Y_i(t_b, M_i(t_b, d_1))]\,. \end{aligned} \tag{18}$$

For any respondents with $M_i(t_b, d_*) = M_i(t_b, d_1)$, the latter two terms of this expression will be equal to each other and so will cancel to 0. Furthermore, because of the above monotonicity assumptions, we know that $M_i(t_b, d_*) \geq M_i(t_b, d_1)$. With this in hand, we can rewrite the decomposition as:

$$\begin{aligned} &NIE_i(t_a, t_b, d_*) - NIE_i(t_a, t_b, d_1) \\ &+ \sum_{j=1}^{J_m} \sum_{k=j+1}^{J_m} \mathbb{I}\left\{M_i(t_b, d_*) = m_k\right\} \mathbb{I}\left\{M_i(t_b, d_1) = m_j\right\} \left[CDE_i(t_a, t_b, m_k) - CDE_i(t_a, t_b, m_j)\right] \end{aligned} \tag{19}$$

Taking expectations of the second part of this expression, we are left with the following version of the imperfect manipulation reference interaction:

$$\begin{aligned} RI^*(t_a, t_b, d_1) = \sum_{j=1}^{J_m} \sum_{k=j+1}^{J_m} \Big(&\mathbb{E}\left[CDE_i(t_a, t_b, m_k) - CDE_i(t_a, t_b, m_j) | M_i(t_b, d_*) = m_k, M_i(t_b, d_1) = m_j\right] \\ &\times \mathbb{P}\left[M_i(t_b, d_*) = m_k, M_i(d_1) = m_j\right]\Big) \end{aligned} \tag{20}$$

Putting this all together, we are left with the following decomposition:

$$TE(t_a, t_b) - CDE^*(t_a, t_b, d_1) = NIE(t_a, t_b, d_*) - NIE(t_a, t_b, d_1) + RI^*(t_a, t_b, d_1) \tag{21}$$

The result given in the main text easily follows by restricting $M_i$ to be binary and $D_i$ to take on three possible values (with a slight adjustment to notation where $d_1$ here corresponds to $d_0$ in the binary case given in the main text). Finally, under the crowd-out assumption, $M_i(t_b, d_1) = M_i(d_1)$, which implies that $NIE_i(t_a, t_b, d_1) = 0$. The decomposition in the main text follows.

# C Additional figures

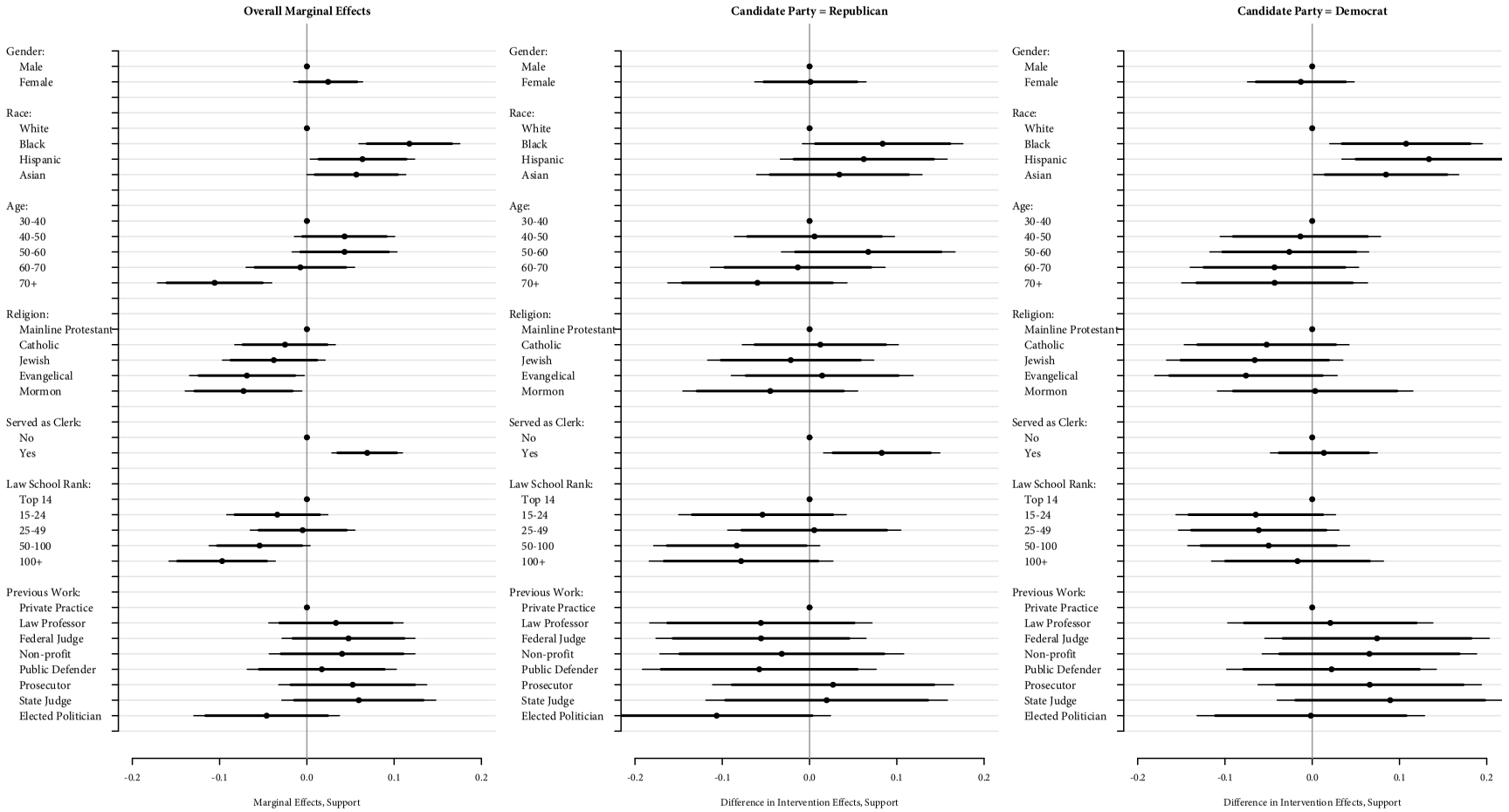


Figure A1: Full results from the conjoint analysis of Sen (2017).