PA

Analyzing Causal Mechanisms in Survey Experiments

Avidit Acharya[1], Matthew Blackwell[2] and Maya Sen[3]

[1] *Assistant Professor, Department of Political Science, Stanford University, Stanford, CA 94305, USA.*
Email: avidit@stanford.edu, URL: http://www.stanford.edu/~avidit
[2] *Assistant Professor, Department of Government, Harvard University, Cambridge, MA 02138, USA.*
Email: mblackwell@gov.harvard.edu, URL: http://www.mattblackwell.org
[3] *Associate Professor, Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA.*
Email: maya_sen@hks.harvard.edu, URL: http://scholar.harvard.edu/msen

Abstract

Researchers investigating causal mechanisms in survey experiments often rely on nonrandomized quantities to isolate the indirect effect of treatment through these variables. Such an approach, however, requires a "selection-on-observables" assumption, which undermines the advantages of a randomized experiment. In this paper, we show what can be learned about casual mechanisms through experimental design alone. We propose a factorial design that provides or withholds information on mediating variables and allows for the identification of the overall average treatment effect and the controlled direct effect of treatment fixing a potential mediator. While this design cannot identify indirect effects on its own, it avoids making the selection-on-observable assumption of the standard mediation approach while providing evidence for a broader understanding of causal mechanisms that encompasses both indirect effects and interactions. We illustrate these approaches via two examples: one on evaluations of US Supreme Court nominees and the other on perceptions of the democratic peace.

Keywords: causal inference, survey experiments, randomized experiments

1 Introduction

Social scientists have increasingly turned to survey experiments to determine the existence of causal effects. Understanding *why* a causal effect exists, however, is also an important goal. For example, do survey respondents have different views about white versus black US presidential candidates primarily because they think white and black politicians tend to belong to different parties? Do respondents support a preemptive American strike against a nondemocratic nuclear power because they tend to assume nondemocracies are threatening to the United States?

Despite the importance of properly assessing such causal mechanisms, a widely used approach has been to observe, rather than manipulate, a potentially mediating variable. In our example of presidential candidates, it would not be unusual for a researcher to ask respondents to speculate about the party of hypothetical black or white presidential candidates and then use that quantity for mediation analyses. In the example of preemptive strikes, a researcher might ask whether respondents believe the rising nuclear power is threatening or not. A downside to this approach, however, is that it "breaks" the experimental design by introducing nonrandomized observational data (e.g., inferred party or speculation about threat), thus raising the specter of omitted variable bias among other issues. This undermines the purpose of using a survey experiment to identify causal effects.

In this paper, we describe a broad set of experimental designs that can, under certain assumptions, speak to causal mechanisms while retaining many of the benefits of experimental

Political Analysis (2018)
vol. 26:357–378
DOI: 10.1017/pan.2018.19

Published
3 August 2018

Corresponding author
Matthew Blackwell

Edited by
R. Michael Alvarez

© The Author(s) 2018. Published by Cambridge University Press on behalf of the Society for Political Methodology.

Many thanks to John Ahlquist, Josh Kertzer, Ryan T. Moore, Paul Testa, and Teppei Yamamoto for helpful feedback. Special thanks to Jessica Weeks and Mike Tomz for sharing their survey instrument with us. Replication data and code can be found in Acharya, Blackwell, and Sen (2018).

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

control. Our approach leverages a different strategy, one that has gained traction among applied researchers, which is to manipulate the information environment. For example, would opinions about white and black presidential candidates change if respondents were provided with the additional cue that both candidates are Republicans? Would opinions about nuclear strikes against nondemocracies persist if we also tell respondents that the country poses little threat to the United States?

As we show, manipulating the informational environment of a survey experiment in this fashion can reveal substantively important patterns. Specifically, we propose a factorial design that randomly assigns the treatment of interest as one factor and the provision or withholding of information about the proposed mediator as another factor. This type of design, proposed in the context of mediation analysis by Imai, Tingley, and Yamamoto (2013), can identify the overall average treatment effect (ATE) 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). In the presidential candidate example, showing respondents additional 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.

By selectively giving respondents information about the mediator, this design can help provide evidence for what role the mediator plays in the mechanism that generates the treatment effect. Our design takes advantage of previous work showing that the difference between the ATE and the controlled direct effect, which we call the *eliminated effect*, can be interpreted as a combination of an indirect (or mediated) effect and a causal interaction (VanderWeele 2015). As we argue, both are components of a causal mechanism, meaning that, when the eliminated effect 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 inherent trade-off 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). Furthermore, these inferences will be robust to assumptions about nonrandomized moderators or mediators, which are violated in a large share of experiments in political science.

A second contribution of this paper 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 information, but rather the respondent's *belief* about that information. If these differ, the average controlled direct effect (ACDE) 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 derive a similar decomposition in this setting. We also show how to interpret results under imperfect manipulation of the mediator. Under these assumptions, we can use the manipulation (what the researcher tells the respondent) rather the mediating variable (what the respondent believes) and still recover a combination of the indirect and interaction effects induced by the manipulation.

Our third contribution is to provide guidance on how intervening on a potential mediator can be (and 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 US Supreme Court and documents how showing the respondents information about the nominee's partisanship reduces the signal conveyed 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

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

Tomz and Weeks (2013) on the theory of the "democratic peace," showing that Americans are less likely to support preemptive strikes against democracies versus nondemocracies. 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 nondemocracies. Importantly, we reach this conclusion without requiring the strong assumptions of the original paper's mediation analysis.

This paper proceeds as follows. We first describe an illustrative example we use throughout, that of a survey experiment assessing public support for US Supreme Court nominees. Next, we introduce the formalism, define the key terms, and explain how our approach differs from others. We then define our three main causal quantities of interest: (1) controlled direct effects, (2) natural-mediator effects, and (3) eliminated effects. We relate these quantities to causal mechanisms under both perfect and imperfect manipulation of the mediator. Furthermore, we show how these quantities apply not just to experiments (and survey experiments in particular), but also more broadly to observational contexts. We then present the two applications, which show that we can identify different quantities of interest depending on the information provided to respondents. We conclude by discussing the implications 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 studying how the public evaluates potential US Supreme Court nominees and whether black and white nominees are evaluated differently. An attractive design for this question would be one that randomly presents respondents with one of two profiles: one with a nominee identified to the respondents as African American and one identified as white. Comparing evaluations of the two profiles would allow the researcher to estimate 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 of the black racial cue could be attributed to racial animus. Or, a negative treatment effect among certain respondents could be attributed to a prior belief that black nominees are more likely to be Democrats (McDermott 1998). Yet another possibility is that a negative treatment effect could be attributed to respondents thinking that white candidates are more likely to have previous judicial experience and are therefore more "qualified." These explanations have different substantive implications: the first mechanism relies on racial prejudice while the second and third point to race as a heuristic for other characteristics.

Manipulating the information environment can help researchers investigate these differing explanations. For example, if the researcher included information about the candidate's partisanship in his experiment (as part of the candidate's profile, perhaps) then he would be able to assess whether the second hypothesis has support. If he included information about the candidate's professional background, 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 manipulating the information environment in survey experiments.

We view the goals of these types of experiments as twofold. First, researchers using such designs want to estimate the baseline causal effects. In our example, is there an effect of nominee race on respondent choice? This is straightforward to assess in an experimental setting, and a large literature in statistics and political science has focused on how to estimate these treatment effects. More complicated is the second goal, which is, given a particular total effect (or marginal component effect, to use the terminology of conjoint experiments), *how* and *why* is there an effect?

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

In our example, our researcher wants to know the mechanism by which the effect came to be—that is, why does race affect a respondent's choice? Although such questions have been of increasing interest, most researchers have proceeded in an *ad hoc* basis or via including observational data back into the analysis, thus undermining the purpose of using an experimental design. Our goal here is to reason more formally about this second goal of investigating mechanisms.

2.1 Mechanisms, mediation, and interaction

We now turn to explaining what we mean by a causal mechanism and how certain experimental designs facilitate their exploration. First, a causal mechanism 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?) Second, in the spirit of counterfactual reasoning, a causal mechanism explains how an intervention or a change in contextual forces could have produced a different result. Thus, 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. We believe that both causal processes and causal interactions speak to the mechanism by which a treatment affects an outcome and that both address the questions we posed above. Both also 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 a causal pathway from treatment to outcome (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 hypothetical nominee's race affects respondents' beliefs about the nominee's partisanship, which in turn affects respondent choice.

Mechanisms as causal interactions

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

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

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" as a shorthand for "potential mechanism variable" 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 respects from existing frameworks. For example, Dafoe, Zhang and Caughey (2017) refer to the changing nature of the treatment effects in the setting that we have in mind as a lack of "informational equivalence." Under their framework, the true treatment effect of a randomized assignment is masked by a respondents' beliefs over other features of the vignette (see also Bansak *et al.* 2017).[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. Below, we do engage with the beliefs of respondents in discussing 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?

Imai, Keele, and Yamamoto (2010) provide assumptions that can identify indirect effects and thus separate them from causal interactions, but these assumptions requires no mediator–outcome confounders (observed or unobserved) other than the treatment and any pretreatment covariates. When the mediator is simply observed and not manipulated, it is difficult to justify the assumption that, to use our example, the race of the nominee is the only confounder for the relationship between inferred partisanship and support for the candidate. Our proposed design sidesteps this concern by randomly assigning the mediator in certain experimental arms. This has the advantage of allowing us to drop any assumptions about mediator–outcome confounders but has the disadvantage of not allowing us to distinguish between causal processes and causal interactions. We explore this trade-off in further detail below. A middle ground between these would be to use our approach as a baseline estimate and then combine the Imai, Keele, and Yamamoto (2010) approach of a sensitivity analysis to explore the range of plausible indirect effects given the experimental design.

Perhaps the 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. 333–336). The approach we take in this paper is a version of this implicit mediation analysis, but we extend the idea 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

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 not be informationally equivalent due to partisanship. That is, respondents might assume that candidates of certain racial or ethnic backgrounds have different partisanships.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

mediator" experiments in Gerber and Green (2012), addressing their concerns about the inability of a researcher to set values of the mediator perfectly.

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 multileveled mediators in the supplemental materials.) In our example, $T_i = 1$ would indicate that a hypothetical Supreme Court nominee was reported to be African American and $T_i = 0$ would indicate that the nominee was reported to be white. The mediator might be partisanship; for example, $M_i = 1$ would indicate that the nominee is identified as a Democrat and $M_i = 0$ 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, $D_i = d_0$ represents informing the subject that the nominee is a Republican (and so M_i should be 0) and $D_i = d_1$ represents informing the subject that the nominee is a Democrat (and so $M_i = 1$).

To define the key quantities of interest, we rely on the potential outcomes framework for causal inference (Neyman 1923; Rubin 1974; Holland 1986). 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 what i infers the party of the nominee to be if only given information 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/ethnicity information *and* partisan information about the hypothetical nominees. For now, we 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, $D_i = d_1$, then the subject believes the candidate is a Democrat, $M_i = 1$. 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

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 a lack of informational equivalence.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

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. If this assumption is violated, then it is difficult to connect the experiment to mechanisms because, in some sense, the manipulation of the mediator and the natural value of the mediator are inherently different concepts and cannot be compared in the same experiment. Thus, this assumption limits the types of mediators that could be studied with our proposed design.

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 $D_i = d_*$, 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 assumption that follows directly from the design of these experiments. 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 natural-mediator arm and that both the treatment and the mediator are randomized in the manipulated-mediator arm. This assumption is substantially weaker than the sequential ignorability assumption of Imai, Keele, and Yamamoto (2010) that justifies the use of standard tools for mediation analysis. Those assumptions require the potential mediator itself to be as good as randomly assigned even though the mediator is only observed, not manipulated, in that setting. This assumption is likely to be false in many settings. In our running example, it would require no unmeasured confounders for the relationship between inferred partisanship of the candidate and their support for the candidate. But there are several likely confounders in this setting, including the policy preferences of the respondent. If we cannot properly measure and control for all of these potential confounders, then the basic assumptions of the mediation approach will be violated and conducting such an analysis is not possible. Essentially, the mediation approach embeds an observational study (of the mediator) into the experimental setting. Our approach avoids these pitfalls by randomly assigning the potential mediator, ensuring that no unmeasured confounders holds by design. A trade-off with our approach is that while the assumptions are much more likely to hold, we cannot identify the same quantities of interest as in a mediation analysis. As we argue below, we believe the quantities of interest we can identify in this setting still provide substantively meaningful evidence on the causal mechanisms at work in the experiment.

This proposed design is an example of a $J_t \times 3$ factorial design, with the first factor being treatment and the second factor being the mediator arm. These experimental designs are common throughout the social sciences, allowing researchers to apply familiar intuitions to this setting. Note that one limitation of this type of design is that when there are a large number of treatment or mediator categories being investigated, the statistical power of the experiment will be constrained and will require large sample sizes. For this reason, we recommend relying on a statistical power analysis to tailor the number of treatment and control categories to a given setting. In many cases, this may require keeping the number of categories to a minimum.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

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 *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 when 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 *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).

If we view this design as a $J_t \times 3$ factorial design, then the ATE and the ACDE represent what Cochran and Cox (1957) call the *simple effects* of treatment at various levels of the mediator-arm factor, D_i.[3]

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 by an amount induced by 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 the 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

3 One benefit of conducting these experiments on a representative survey is that it provides a firmer basis for generalization of both the ATE and the ACDE. That is, the "natural mediator" in the natural-mediator arm has the interpretation of being an unbiased estimate of the natural value of the mediator in the population. In this way, representative surveys protect this design from sample-induced biases when generalizing our inferences to the population. We thank an anonymous reviewer for highlighting this issue.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

(for the mediator). Since we never observe both of these states at once, identification of the natural indirect effect will often require strong and perhaps unrealistic assumptions. As the name implies, the quantity 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.[4] As above, we define the *average natural indirect effect* (ANIE) to be $NIE(t_a, t_b) = \mathbb{E}[NIE_i(t_a, t_b)]$.

Reference interactions

To capture the interaction between T_i and M_i, we introduce the *reference interaction* (VanderWeele 2014), which is the difference between the direct effect under the natural value of the mediator under t_b, or $M_i(t_b, d_*)$ and the controlled direct effect for the reference category m:

$$RI_i(t_a, t_b, m) = [Y_i(t_a, M_i(t_b, d_*)) - Y_i(t_b, M_i(t_b, d_*))] - [Y_i(t_a, m) - Y_i(t_b, m)]. \quad (4)$$

In our example, the reference interaction would compare the direct effect 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, $RI(t_a, t_b, m) = \mathbb{E}[RI_i(t_a, t_b, m)]$. 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 Republican), then this quantity is the average difference between the direct effect of race under inferred partisanship and the direct effect of race for a Republican profile. The ARIE provides a summary measure of how the ACDE varies across units due to natural 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 at the individual level, or (2) the natural value of the mediator under t_b is always equal to m (e.g., all white profiles are inferred to be Republicans). In both cases there is no interaction, either because the treatment effects or the natural value of the mediator does not vary. This quantity may be equal to zero if there are exact cancelations in the interactions across the population, but this is both rare and dependent on the baseline category, m.

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, leading 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:

$$NME_i(t, m) = Y_i(t) - Y_i(t, m) = Y_i(t, M_i(t, d_*)) - Y_i(t, m). \quad (5)$$

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

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

This quantity is 0 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. Respondents who infer the partisanship of the hypothetical candidate to be a Democrat will have an NME of zero since, for them, their natural value is equal to the baseline value, $M_i(t, d_*) = m$. 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 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 intervention involves manipulating the information environment.

Eliminated effect

The nature of a causal mechanism is about how much of a treatment effect is explained by a particular mediator. To measure this, we define the *eliminated effect* as the difference between the overall ATE and the ACDE :[5]

$$\Delta_i(t_a, t_b, m) = \underbrace{[Y_i(t_a, M_i(t_a, d_*)) - Y_i(t_b, M_i(t_b, d_*))]}_{\text{total effect}} - \underbrace{[Y_i(t_a, m) - Y_i(t_b, m)]}_{\text{controlled direct effect}}. \tag{6}$$

In the context of the Supreme Court nominee example, this quantity 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. Thus, the eliminated effect represents the amount of the total effect that is eliminated by setting party to a particular value.

Another way to derive and understand the eliminated effect is as the difference between two natural-mediator effects. In particular, it is simple to show that the eliminated effect can be written as:

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

One 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. From this point of view, the eliminated effect is 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. As above, we will focus on the

5 We take this naming from Robins and Greenland (1992), who 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."

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

Table 1. Representation of the different effects described in the proposed design. The interior cells show what the average outcome of the experimental arm identifies. The margins show what effects correspond to the difference of the quantities in the rows and columns. The eliminated effect, Δ, is the difference between these differences. For clarity, we only include one manipulated-mediator arm.

	Treatment (T_i)		
Mediator arm (D_i)	**Black profile (t_a)**	**White profile (t_b)**	**Difference**
Inferred-party arm (d_*)	$\mathbb{E}[Y_i(\texttt{black})]$	$\mathbb{E}[Y_i(\texttt{white})]$	$TE(\texttt{black}, \texttt{white})$
Manipulated-party arm (d_0)	$\mathbb{E}[Y_i(\texttt{black}, \texttt{dem})]$	$\mathbb{E}[Y_i(\texttt{white}, \texttt{dem})]$	$ACDE(\texttt{black}, \texttt{white}, \texttt{dem})$
Difference	$ANME(\texttt{black}, \texttt{dem})$	$ANME(\texttt{white}, \texttt{dem})$	$\Delta(\texttt{black}, \texttt{white}, \texttt{dem})$

average of these eliminated effects,

$$\Delta(t_a, t_b, m) = \mathbb{E}[TE_i(t_a, t_b) - CDE_i(t_a, t_b, m)] = TE(t_a, t_b) - CDE(t_a, t_b, m),$$

which is simply the difference in the ATE and the ACDE at the manipulated level of the mediator m.

Table 1 summarizes both the factorial nature of our proposed design and the various quantities of interest that can be identified from this design via contrasting various experimental arms. This table also highlights how the eliminated effect is a differences-in-differences quantity.

3.2 How eliminated effects help us understand causal mechanisms

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

$$\Delta_i(t_a, t_b, m) = \underbrace{NIE_i(t_a, t_b)}_{\text{indirect effect}} + \underbrace{RI_i(t_a, t_b, m)}_{\text{interaction effect}}. \tag{8}$$

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 eliminated effect 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 eliminated effect when party is fixed to "Democrat" 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. 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 Democrat. In some contexts, this latter condition may be plausible. For example, given that few African Americans identify as Republicans, assuming 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 eliminated effect 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 eliminated effect is impossible. To do so would require

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

stronger assumptions such as 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 eliminated effect 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, as discussed above, 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 the 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

Thus far, we have assumed that 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 monotonicity assumption that puts structure on the manipulations:

ASSUMPTION 3 (Monotonicity). For all $t \in \mathcal{T}$, $M_i(t, d_0) \leq M_i(t, d_*) \leq M_i(t, d_1)$.

Monotonicity states that providing information does not have perverse effects. For example, suppose that d_1 here refers to "Democrat" and d_0 corresponds to "Republican," where treatment is still the race of the candidate. This assumption rules out pathological cases where under no manipulation the respondent believes a candidate is a Democrat ($M_i(t, d_*) = 1$), but when told that the candidate is a Democrat would believe that the candidate is a Republican ($M_i(t, d_1) = 0$). 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:

$$CDE^*(t_a, t_b, d_0) = \mathbb{E}[Y_i(t_a, M_i(d_0)) - Y_i(t_b, M_i(d_0))]. \tag{9}$$

This is the estimand that would be identified in the manipulated-mediator arm under imperfect manipulation, so long as the exclusion restriction (Assumption 1) and randomization (Assumption 2) hold. We can also define similarly modified versions of the eliminated effect, $\Delta^*(t_a, t_b, d_0) = TE(t_a, t_b) - CDE^*(t_a, t_b, d_0)$. These effects are now defined in terms of the experimental-arm

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

manipulation rather than the mediator directly. We also need to define versions of the indirect effect and reference interaction in this setting. First, under imperfect manipulation there may be indirect effects even in the manipulated-mediator arms, so that $NIE(t_a, t_b, d_0) = \mathbb{E}[Y_i(t_a, M_i(t_a, d_0)) - Y_i(t_a, M_i(t_b, d_0))]$. Under perfect manipulation, this quantity would be zero since the manipulation, d_0, would completely determine the mediator. Second, we can redefine the reference interaction to be the manipulation-induced reference interaction:

$$RI^*(t_a, t_b, d_0) = \mathbb{E}[Y_i(t_a, M_i(t_b, d_*)) - Y_i(t_b, M_i(t_b, d_*))] - \mathbb{E}[Y_i(t_a, M_i(t_b, d_0)) - Y_i(t_b, M_i(t_b, d_0))].$$

This interaction represents the change in the direct effect of treatment for those units that update their beliefs when provided information d_0. This group would believe that a white nominee is a Democrat when not provided with partisanship, $M_i(t_b, d_*) = 1$, but would change their mind if told that the nominee is a Republican, $M_i(t_b, d_0) = 0$. Then, we show in the supplemental materials that the following decomposition holds:

$$\Delta^*(t_a, t_b, d_0) = \underbrace{NIE(t_a, t_b) - NIE(t_a, t_b, d_0)}_{\text{manipulation-induced indirect effect}} + \underbrace{RI^*(t_a, t_b, d_0)}_{\text{manipulated-induced interaction}}. \quad (10)$$

This decomposition shows that the eliminated effect (at level d_0) under imperfect manipulation is the sum of two components. First is the difference between the indirect effect in the natural-mediator arm and the indirect effect in the manipulated-mediator arm. We call this the *manipulation-induced* indirect effect because it is the portion of the total indirect effect of M_i that is just due to the manipulation. Second is the manipulation-induced reference interaction. This modified interaction effect is the difference in the CDE induced by respondents who update their beliefs in response to manipulation. Thus, we can interpret the eliminated effect with imperfect manipulation as the sum of the indirect effects and interactions *due the manipulation alone*. Under perfect manipulation of the mediator, these two quantities become the usual natural indirect effect and reference interaction.

This analysis highlights how the strength of the manipulation is important when interpreting the eliminated effect. It is straightforward to show that both the manipulation-induced indirect effect and reference interaction will be zero when the manipulation has no effect on the mediator, or $M_i(t, d_*) = M_i(t, d_0)$. This makes sense in the context of the exclusion restriction (Assumption 1), since it requires the manipulation to have no effect other than through the mediator. If it has no or a weak effect on the mediator, the eliminated effect will be close to zero. Thus, when the ATE and the ACDE are similar, it could be due to the mediator not being part of a causal mechanism or it could be because the manipulation of the mediator is weak. On the other hand, this indicates that imperfect manipulation will generally lead to underestimates of the true eliminated effect, meaning that our estimates of causal mechanisms will be conservative in this setting.

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 conjoint 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 carryover effects, simple difference-in-means estimators that aggregate across

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

respondents are unbiased for what they call the *average marginal component effect* (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 eliminated effect 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 posttreatment bias

When thinking about variables possibly affected by the treatment of interest, a common threat to inference is *posttreatment bias* (Rosenbaum 1984). Posttreatment bias can occur when conditioning on a variable that is affected by the treatment (making it "posttreatment"). It is useful to partition this bias into two different types that are often conflated. First, conditioning on a posttreatment 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 posttreatment 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 problems. The first type of posttreatment bias is actually our target of estimation here—the difference between the ATE and ACDE. And, because the studies we consider here are ones that experimentally manipulate the mediator, selection bias does not arise here. In observational studies, on the other hand, posttreatment 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 ATE and the controlled direct effect are both identified in an observational study, the decomposition in (6) implies that we can also identify the eliminated effect. 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 eliminated effect 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 above decomposition of the ATE suggests that the eliminated effect 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 plow use affects contemporary attitudes toward 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 plow use on contemporary attitudes toward women, the eliminated effect can be thought of in the following way. First consider the overall effect of plow use on contemporary attitudes toward women. Then consider intervening on countries with different levels of plow use so that they have the same level of the contemporary income. If the effect of plow use disappears after this intervention, we might interpret this as evidence that contemporary income helps explain the effect of historical plow use, either through mediation or interaction. However, if they are the same, we might interpret it as evidence that

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

income is not part of a causal mechanism. 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, the eliminated effect under imperfect manipulation of the mediator is identified as:

$$\begin{aligned}\Delta^*(t_a, t_b, d_m) = &[\mathbb{E}[Y|T_i = t_a, D_i = d_*] - \mathbb{E}[Y|T_i = t_a, D_i = d_m]] \\ &- [\mathbb{E}[Y|T_i = t_b, D_i = d_*] - \mathbb{E}[Y|T_i = t_b, D_i = d_m]].\end{aligned} \tag{11}$$

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 eliminated effect 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:

$$\widehat{\mathbb{E}}[Y_i|T_i = t_a, D_i = d_*] = \frac{\sum_{i=1}^{N} Y_i \mathbb{I}\{T_i = t_a, D_i = d_*\}}{\sum_{i=1}^{N} \mathbb{I}\{T_i = t_a, D_i = d_*\}}. \tag{12}$$

Replacing each of the expectations in (11) 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, 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 to the natural-mediator arm and the manipulated-mediator arm with mediator value m ($D_i \in \{d_*, d_m\}$) 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 eliminated effect, $\Delta(t_a, t_b, m)$. 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 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) show 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 eliminated effect 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).

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

5 Experimental Analysis of Direct Effects and Mechanisms

5.1 Study #1: conjoint experiment for nominees to the US Supreme Court

Our first application is an example from Sen (2017) on how the public views nominees to the US Supreme Court. This experiment provides an attractive illustration since the true ideological leanings of Supreme Court nominees is often noisily conveyed to the public. Half of the 1,650 respondents in the study were randomly assigned to see 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.[6]

This experimental design matches our setting well. In the absence of partisan cues, racial information contained in the profiles may activate 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 total effect of the "black" racial cue (marginalizing over other attributes in the conjoint) should be positive for Democratic respondents. However, introducing information about 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 fixing partisanship of the nominee. From these two experimental arms, we can estimate the eliminated effect, $\Delta(t_a, t_b, m)$. If this quantity were zero, then we could conclude that partisanship plays little role in the effect of race on support. Indeed, this would mean that giving information about the partisanship of the nominee had no impact on the effect of race on support. If this quantity were positive, it would indicate that some portion of the positive effect of race is due to inferred partisanship, either through indirect effects or interactions. If the eliminated effect is equal to the total effect, then this implies that the ACDE of race fixing partisanship is equal to zero and that any effect of race in the natural-mediator arm is due to inferred partisanship.

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 information 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 AMCEs for each race category from the natural-mediator arm, then estimate the ACDEs from one of the manipulated-mediator arms, and then use these two quantities to estimate the eliminated effect. In other words, the eliminated effect is the difference between the effect of a black profile (versus a white profile) under no party information and the same effect when party is set to Republican *or* Democrat.[7]

Figure 1 shows the results for the effects of nominee race, with the total AMCEs in the left panel and the eliminated effects 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 marginal effects, we can see that Democratic respondents are more like to support minority nominees. (In supplemental material 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

6 This study was conducted in December of 2013 with a nonprobability sample recruited by Survey Sampling International matched to the US adult population based on age, gender, race, and geography. Replication data and code can be found in Acharya, Blackwell, and Sen (2018). See also Kirkland and Coppock (2017) for a similar design in a slightly different context.

7 Note that there are two possible ACDEs, one for Republican (noncopartisan) and Democratic (copartisan) profiles and so there are two possible eliminated effects corresponding to each of these.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

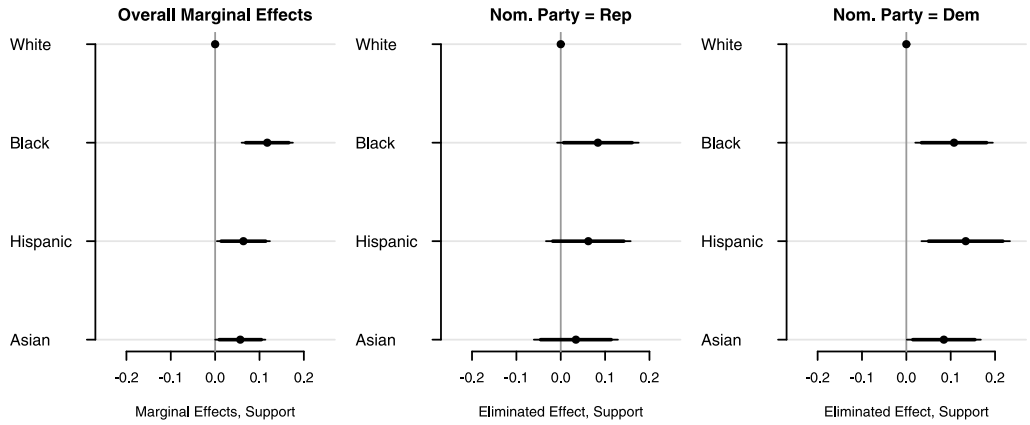

Figure 1. Average marginal effects of nominee race on support for the nominee (left panel) and eliminated effects for nominee partisanship as a mediating variable (middle and right panels). The eliminated effect in the middle panel has the partisanship set to "Republican" and the eliminated effect in 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.

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 eliminated effect tell us this exactly. In the right two panels, we show this quantity 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 eliminated effects 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 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 identified, so long as the partisan affiliation manipulation is monotonic for beliefs about partisanship 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. A positive indirect effect would indicate that race impacts inferred partisanship and that this changes support for the candidate. A positive reference interaction would imply that the direct effect of race under the inferred partisanship of a white nominee is greater than the direct effect under the manipulated partisanship. That is, inferring partisanship caused a higher direct effect of race. 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 ignorability 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. For reasons discussed above, this may be

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

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 nondemocracies 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 was a democracy, (2) whether the country had a military alliance with the United States, and (3) whether the country had a high level of trade with the United States. The authors then asked a set of follow-up questions to assess perceptions of threat, such as whether or not the respondent thought it likely that the country would initial a nuclear strike against a neighbor or the United States. Of particular interest to us is that the authors then used the answers to these threat questions in a mediation analysis to explore how perceptions of threat may mediate the effect of democracy on support for a strike. However, the mediation analysis requires that there be no unmeasured confounders between perceptions of threat and support for an attack. Because the question on perception of threat was simply measured and not randomized, the "no unmeasured confounders" assumption could be unreasonably strong.

To explore these mechanisms, we fielded a replication and extension of this study on June 9th, 2016. We recruited a sample of 1,247 respondents from Mechanical Turk who took the online survey through Qualtrics.[8] In this study, 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.[9] We then randomly manipulated one additional treatment condition. Half of the respondents were randomly assigned to the experimental design exactly as it was in Tomz and Weeks (2013), with no information given about the threat that the hypothetical country poses. Half of respondents were randomly assigned to the manipulated-mediator arm, which contained additional information about the threat in the vignette: "The country has stated that it is seeking nuclear weapons to aid in a conflict with another country in the region."[10] 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 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 nondemocracy (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

8 The sample was restricted to adults aged 18 or older residing in the United States. Respondents were told this was a five-minute survey about international affairs and were offered $1.25 for their participation. We additionally collected some demographic information from respondents, but do not use these in the analyses below. We present the full survey instrument in the supplemental materials. Replication data and code can be found in Acharya, Blackwell, and Sen (2018).

9 Tomz and Weeks provided us with the exact survey instrument used for their experiment and we followed that instrument exactly with the exception of the manipulated mediator.

10 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.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

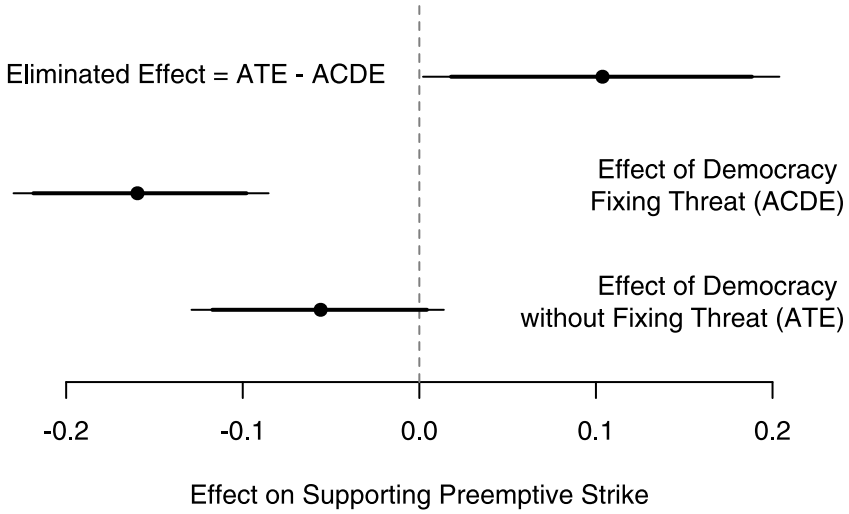

Figure 2. Results from the replication of Tomz and Weeks (2013). Data from a Mechanical Turk survey experiment ($N = 1247$). Bootstrap 95% (thin line) and 90% (thick line) confidence intervals are based on 5,000 bootstrap replications.

unusual instance in the sense that the ACDE is actually larger in magnitude than the ATE.[11] This difference leads to an overall positive eliminated effect, which, as above, can be thought of as the combination of the natural indirect effect and the reference interaction.

This positive eliminated effect stands in contrast to Tomz and Weeks (2013), who found a negative indirect effect of democracy through potential threat. There are two possible reasons for this. First, the indirect effect might indeed be negative in our experiment as well, but the reference interaction is so positive that the overall result is a positive eliminated effect. A positive reference interaction in this case would indicate that the direct effect of democracy at the inferred level of threat for an autocratic profile is lower in magnitude (i.e., less negative) than the direct effect under the manipulated level of threat in this study. The inferred level of threat under autocracy might be quite high if, for example, respondents inferred the hypothetical country to be North Korea, a country that has threatened the United States in the past. The level of threat given in the manipulated-mediator arm was that the hypothetical country was seeking nuclear weapons as part of a dispute in their region. This level of threat would be lower than the inferred threat of North Korea since it does not directly threaten the United States. Then, if higher levels of threat tend to lower the effect of democracy, then the direct effect under the inferred mediator would be closer to zero than the direct effect under the manipulated mediator and leave us with a positive reference interaction.

A possible second explanation for the divergence with Tomz and Weeks (2013) is that the natural indirect effect might be positive or nonexistent in both studies, but a violation of the usual mediation assumptions in Tomz and Weeks (2013) study could have led to a biased estimate of the indirect effect. In that study, Tomz and Weeks measured the perceived level of threat after respondents read the profile. The usual sequential ignorability assumption of mediation analysis (Imai, Keele, and Yamamoto 2010) would require there to be no unmeasured confounders between perceived threat and support for a preemptive strike other than what is in the profile and any pretreatment covariates. This would be violated if, for example, some respondents had

11 From Figure 2, it appears that the difference between the ATE and the ACDE may not be statistically significant since their confidence intervals overlap. This appears in contrast to the confidence intervals for the eliminated effect that do not include zero. This apparent discrepancy is due to the covariance between the two estimates leading to confidence interval widths for the eliminated effect that are less than the sum of the widths of the ATE and ACDE. This result holds with standard variance estimators for OLS, Huber-White "robust" standard errors, or the bootstrapping approach used here.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

higher levels of anxiety that affected both perceptions of threat and support for preemptive strikes. Unfortunately, this very standard experimental design for mediation essentially creates an observational study within the experiment. If these strong assumptions do not hold, then the estimates of the indirect effect in Tomz and Weeks (2013) could be biased. Our experimental design does not rely on this assumption and, thus, would not suffer from this type of bias. In this way, the positive eliminated effect could be giving us evidence of a positive or null indirect effect and still be consistent with the results of Tomz and Weeks (2013).

These two explanations are not mutually exclusive and could work together to produce the large, positive eliminated effect 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.[12]

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 eliminated 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 US 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 eliminated effect 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 eliminated effect 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

12 There is, of course, another possibility: that the samples generated by the two experiments were different enough to lead to different ATE and ACDE estimates. This might be plausible if the Mechanical Turk pool has changed dramatically over the last few years or if the incentives to participate between the two studies were different. We think that this is less of an issue in this case given the similarity of the ATE estimates being similar across the two studies.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

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.

Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2018.19.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. Explaining causal findings without bias: detecting and assessing direct effects. *American Political Science Review* 110(3):512–529.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2018. Replication data for: analyzing causal mechanisms in survey experiments. doi:10.7910/DVN/KHE44F, Harvard Dataverse, V1, UNF:6:VaeMnsesFmBVwg98eK/heA==.

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.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2017. Beyond the breaking point? Survey satisficing in conjoint experiments. Stanford University Graduate School of Business Research Paper No. 17-33; MIT Political Science Department Research Paper No. 2017-16.

Cochran, William G., and Gertrude M. Cox. 1957. *Experimental designs*. New York: John Wiley & Sons.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2017. Information equivalence in survey experiments. *Political Analysis*, forthcoming.

Gerber, Alan S., and Donald P. Green. 2012. *Field experiments: design, analysis, and interpretation*. New York: 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(4):765–789.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* 25(1):51–71.

Kirkland, Patricia, and Alexander Coppock. 2017. Candidate choice without party labels: new insights from conjoint survey experiments. *Political Behavior*. Epub ahead of print, doi:10.1007/s11109-017-9414-8.

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: 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.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 6(5):688–701.
Sen, Maya. 2017. How political signals affect public support for judicial nominations: evidence from a conjoint experiment. *Political Research Quarterly* 70(2):374–393.
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):749–761.
VanderWeele, Tyler J. 2015. *Explanation in causal inference: methods for mediation and interaction*. New York: Oxford University Press.
VanderWeele, Tyler J., and Eric J. Tchetgen Tchetgen. 2014. Attributing effects to interactions. *Epidemiology* 25(5):711–722.

https://doi.org/10.1017/pan.2018.19 Published online by Cambridge University Press