Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments*

Kosuke Imai† Teppei Yamamoto‡

First Draft: May 17, 2011
This Draft: January 10, 2012

Abstract

Social scientists are often interested in testing multiple causal mechanisms through which a treatment affects outcomes. A predominant approach has been to use linear structural equation models and examine the statistical significance of corresponding path coefficients. However, this approach implicitly assumes that the multiple mechanisms are causally independent of one another. In this paper, we consider a set of alternative assumptions that are sufficient to identify the average causal mediation effects when multiple, causally related mediators exist. We develop a new sensitivity analysis for examining the robustness of empirical findings to the potential violation of a key identification assumption. We apply the proposed methods to three political psychology experiments which examine alternative causal pathways between media framing and public opinion. Our analysis reveals that the validity of original conclusions is highly reliant on the assumed independence of alternative causal mechanisms, highlighting the importance of proposed sensitivity analysis.

Key Words: causal mediation analysis, experimental designs, linear structural equation modeling, multiple mediators, sensitivity analysis

*We are grateful to Ted Brader, Jamie Druckman, and Rune Slothuus for sharing their data with us. We thank Dustin Tingley and Mike Tomz for useful discussions that motivated this paper. Adam Glynn and Tyler VanderWeele provided helpful suggestions. Financial support from the National Science Foundation (SES–0918968) is acknowledged. An earlier version of this paper was circulated under the title of "Sensitivity Analysis for Causal Mediation Effects under Alternative Exogeneity Conditions."

†Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

‡Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: teppei@mit.edu, URL: http://web.mit.edu/teppei/www

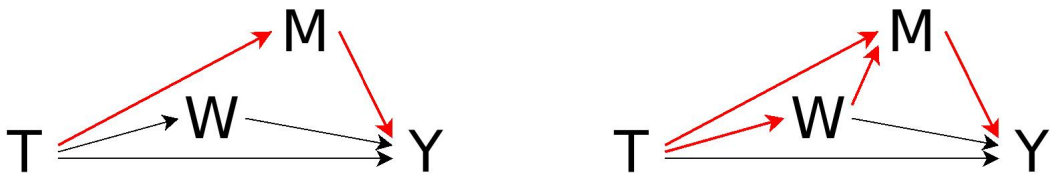
1 Introduction

The identification of causal mechanisms is an important goal of empirical social science research. Researchers are often not only interested in the question of whether a particular causal variable of interest affects outcomes, but they also wish to understand the mechanisms through which such causal effects arise. Causal mediation analysis represents a formal statistical framework that can be used to study causal mechanisms. For example, Imai et al. (2011) identify a set of commonly invoked assumptions under which such an analysis can be justified, develop general estimation algorithms, and propose sensitivity analysis and new research design strategies. Indeed, there exists a fast growing methodological literature about how to study causal mechanisms through the use of statistical methods (see Robins and Greenland (1992); Pearl (2001); Petersen et al. (2006); VanderWeele (2009); Imai et al. (2010c); Glynn (2012) among many others). With such methodological advances, researchers can now base their mediation analysis on a rigorous statistical foundation.

Nevertheless, an important deficiency of the existing methodologies is their limited ability to handle multiple causal mechanisms of interest (see e.g., Bullock et al., 2010; Imai et al., 2011). Given this current state of the literature, the prevailing practice among applied researchers is to assume, often implicitly, that no causal relationship exists among these alternative mechanisms. In many applications, however, such an assumption is not credible because competing theoretical explanations are often closely tied to each other.

In this paper, we consider the identification and sensitivity analysis of multiple causal mechanisms by extending causal mediation analysis to the cases involving several mediators that are causally related to each other. Figure 1 presents two causal diagrams with multiple causal mechanisms. They represent the specific scenarios studied in this paper where the treatment variable T could affect the outcome Y in three ways: through the main mediator of interest M (red arrows), through alternative mediators W, and directly. The critical difference between Figures 1(a) and 1(b) is that in the former M and W are assumed to have no causal relationship with one another while in the latter W may possibly affect M (but not the other way around). We first show that the standard mediation analysis assumes the causal independence between multiple mediators. We then develop a set of statistical methods to relax this assumption.

In Section 2, we introduce three experimental studies from the political psychology literature on framing effects, which we use as illustrative examples throughout this paper. All of these studies investigate how issue frames affect opinions and behavior through multiple causal pathways, such as changes in perceived issue importance or belief content. To identify these multiple mechanisms, they



(a) (b)

Figure 1: Causal Diagrams with Multiple Causal Mechanisms. In Panel (a), the treatment variable T can affect the outcome variable Y either through the mediator of interest M (red arrows), through alternative mediators W, or directly (solid bottom arrow). In addition, the causal relationship between M and W is assumed to be non-existent. Panel (b) represents an alternative scenario where W can affect Y either directly or through M, thereby allowing for the potential causal relationship between M and W. The quantity of interest for the proposed methods is the average causal mediation effect with respect to M, which is represented by the red arrows connecting T and Y through M in both panels.

rely on a traditional path-analytic method which, as we show, implicitly assumes the absence of causal relationship between the corresponding mediators. This assumption is problematic from a theoretical point of view because, for example, framing may also alter the perceived importance of the issue by changing the factual content of beliefs (Miller, 2007).

In Section 3, we use a formal statistical framework of causal inference and show that the standard causal mediation analysis, including the ones conducted in these framing studies, assumes the causal independence between multiple causal mechanisms. After reviewing the assumptions required for the identification of independent multiple mechanisms, we reanalyze the data from the framing experiments under these assumptions. The results suggest that the empirical conclusions in the original studies are largely valid so long as these mechanisms are independent of each other, although statistical significance appears to be somewhat lessened for some of the studies.

In Section 5, we develop new statistical methods that allow for the existence of multiple causal mechanisms that are causally related to each other. Our methods are formulated within a framework of familiar linear regression models, and yet the models are much more flexible than standard regression models because every coefficient is allowed to vary across individual observations in an arbitrary fashion. This flexible varying coefficient model encompasses a large class of realistic statistical models that can be applied to various social science research projects while maintaining the simplicity and interpretability of the linear model.

Under this framework of multiple mechanisms, we first consider the identification assumptions originally introduced by Robins (2003) that allow for the presence of competing mechanisms. Robins assumes that the treatment assignment is exogenous given a set of observed pre-treatment covariates and that the mediator of interest is also exogenous conditional on observed pre-treatment and post-treatment covariates. Thus, the setup allows for the existence of alternative mediators that confound the relationship between the main mediator of interest and the outcome. Under this circumstance,

Robins shows that the assumption of no interaction effect between the treatment and the mediator (i.e., the causal effect of the mediator on the outcome does not depend on the treatment status) is sufficient for identifying the causal mechanism of interest.

As previously noted by many researchers (e.g., Petersen et al., 2006; Imai et al., 2012), however, the assumption of no treatment-mediator interaction is often too strong in empirical applications. For example, in framing experiments, the effect of the perceived issue importance on opinions may well depend on which frame was initially given. Therefore, we relax this key identification assumption by developing a new sensitivity analysis that assesses the robustness of empirical results to the potential violation of this key identification assumption. Our sensitivity analysis also directly addresses "the product and difference fallacies" pointed out by Glynn (2012) because his model is a special case of our model. We illustrate the proposed methods by applying them to the framing experiments from the political psychology literature. Our analysis reveals that the validity of original findings is highly reliant on the assumed independence of alternative causal mechanisms, implying the essential role of identification and sensitivity analyses in the study of multiple mediators.

Causal mediation analysis like many other tools of causal inference relies on untestable assumptions. Thus, it is essential for applied researchers to examine the robustness of empirical findings to the violation of key identification assumptions. While there exist sensitivity analysis tools in the literature, many of them deal with *pre-treatment* confounders and essentially assume the absence of causally related alternative mediators (e.g., Hafeman, 2008; Imai et al., 2010a,c; VanderWeele, 2010). In contrast to these previous studies, the sensitivity analysis proposed in this paper addresses the possible existence of *post-treatment* confounders. Sensitivity analyses that can handle post-treatment confounders are just beginning to appear in the methodological literature on causal mediation (e.g., Albert and Nelson, 2011; Tchetgen Tchetgen and Shpitser, 2011).[1]

Nevertheless, one limitation of the proposed approach is that, even though it addresses the potential violation of the no treatment-mediator interaction assumption, it still hinges on other assumptions such as the exogeneity of the mediator. To address this issue, in Section 7, we consider the extensions of the proposed method to the new experimental designs recently developed by Imai et al. (2012). In particular, we consider the "parallel design" where the sample is randomly divided into two groups and separate randomized experiments are conducted in parallel for those groups.[2] We also consider

[1] Albert and Nelson (2011) assume a stronger version of exogeneity which is by itself sufficient to identify the average causal mediation effects (e.g., the sequential ignorability of Imai et al., 2010c) in order to consider the estimation of other path-specific effects. The objective of Tchetgen Tchetgen and Shpitser (2011)'s method is similar to ours. Their model is more general than ours but it relies on more complex sensitivity parameters than our sensitivity analysis.

[2] In one experiment, researchers randomize the treatment and observe the values of the mediator and the outcome. In the other experiment, both the treatment and mediating variables are randomized and subsequently the values of the outcome variable are recorded.

the "parallel encouragement design", a natural generalization of the parallel design for the situation where the direct manipulation of the mediator is difficult. Both experimental designs allow researchers to relax the exogeneity assumption with respect to the treatment and mediator. In addition, these new experimental designs can serve as templates for observational studies to guide researchers to credible research designs for identifying causal mechanisms (see Imai et al., 2011, for some examples).

Finally, Section 8 offers concluding remarks and some suggestions for applied researchers who wish to study multiple causal mechanisms. All of our proposed methods can be implemented through the open-source software program `mediation` (Imai et al., 2010b), which is freely available at the Comprehensive R Archive Network (CRAN http://cran.r-project.org/package=mediation).

2 Framing Experiments in Political Psychology

In this section, we introduce three empirical studies of framing effects, which serve as examples throughout the paper. In political psychology, scholars are interested in whether and how the framing of political issues in mass media and elite communications affects citizens' political opinion and behavior. Psychological theory suggests that issue framing, or a presenter's deliberate emphasis on certain aspects of an issue, may affect how individuals perceive the issue and change their attitudes and behavior (Tversky and Kahneman, 1981).

If citizens are prone to such cognitive biases when interpreting media contents, political elites may be able to influence, or even manipulate, the public opinion by carefully choosing the languages they use in their communication through mass media (Zaller, 1992). While early studies focused on identifying the issue areas in which such framing effects manifest (e.g. Kinder and Sanders, 1990; Iyengar, 1991; Nelson and Kinder, 1996), recent studies address the question of the mechanisms through which framing affects public opinion and political behavior (e.g. Nelson et al., 1997; Callaghan and Schnell, 2005; Chong and Druckman, 2007).

Below, we briefly describe three experimental studies that are aimed at the identification of such mechanisms. Aside from their prominence in the literature, these studies all explicitly examine more than one causal mechanism by measuring multiple mediators corresponding to those competing possible pathways. In each of these studies, the authors implicitly assume that the multiple causal pathways under study are independent of one another. As we discuss below, however, this assumption is theoretically implausible and statistically problematic, for its violation can lead to a substantial bias in the estimated importance of causal mechanisms.

2.1 Druckman and Nelson (2003)

One of the most important debates in the framing effects literature concerns whether issue framing affects citizens' opinions by shifting the perceived importance of the issue (hereafter the "importance" mechanism) or changing the content of their belief about the issue (hereafter the "content" mechanism). As part of their experimental study on the interaction between issue framing and interpersonal conversations, Druckman and Nelson (2003) examine this question using a path-analytic approach. First, they randomly assign each of their 261 study participants to one of the two conditions. In one condition, the subject is asked to read an article on a proposed campaign finance reform which emphasizes its possible violation of free speech. In the other condition, the assigned article emphasizes the potential of campaign finance reform to limit special interests. Then, after additionally randomizing whether the participants will engage in discussion, the authors measured the two mediators; the participants' perceived importance of free speech and special interests as well as their belief about the impact of the proposed reform on these items. Finally, the authors measured the outcome variable, the overall level of support for the proposed campaign finance reform. The substantive question of interest for the original authors is whether the effect of the frames on the support levels are mediated by the importance mechanism or the content mechanism.

2.2 Slothuus (2008)

Following up on Druckman and Nelson (2003), Slothuus (2008) conducted a randomized experiment to analyze the above two mechanisms of framing effects. Using a sample of 408 Danish students, the author examined how two versions of a newspaper article on a social welfare reform bill – one emphasizing the reform's supposed positive effect on job creation and the other focusing on its negative impact on low-income population – affect differently the participants' opinion about the reform. After randomly assigning participants to either the "job frame" or the "poor frame," the author measured the two mediators, i.e., the importance and content of issue-related considerations, by asking a series of five-point scale questions. Finally, the outcome variable was measured by asking the participants whether and to what extent they agree or disagree with the proposed welfare reform. Similar to the previous study, the key substantive question of interest is whether the framing effects transmit through the importance mechanism or the content mechanism.

2.3 Brader, Valentino and Suhay (2008)

The third study we analyze also investigates the causal mechanisms underlying framing effects but focuses on the role of emotions as opposed to more conscious beliefs about the issue. Brader et al. (2008) report the results of their randomized experiment on the framing of immigration policy. As part

of a nationally representative survey of 354 white non-Latino adults, they randomly assigned different versions of a mock *New York Times* article about immigration by varying the origin of the featured immigrant as well as the tone of the story. The article featured either a European or Latino immigrant and emphasized either the positive or negative consequences of increased immigration. After the treatment, the authors measured the participants' belief about the likely negative impact of immigration (hereafter the "perceived harm" mechanism) as well as their emotions by asking how they feel about increased immigration (hereafter the "anxiety" mechanism). Finally, the authors recorded the participants' opinions and behavioral reactions to increase in immigration. The main goal is to identify the mechanism through which the framing effects of the news stories operated.

3 Identification of Independent Multiple Mechanisms

In this section, we use the formal statistical framework of causal inference and show that the standard causal mediation analysis commonly used in empirical studies (including our running examples introduced in the previous section) implicitly assumes the independence of competing causal mechanisms. We then analyze the framing experiments described above under this independence assumption.

3.1 Causal Mediation Analysis with a Single Mediator: A Review

We begin by briefly reviewing the standard causal mediation analysis with a single mediator (see Imai et al., 2011, for a more detailed explanation). Suppose that we have a simple random sample of size n from a population of interest $\mathcal{P}$. Let T_i be a binary treatment variable, which equals 1 if unit i receives the treatment and is equal to 0 otherwise. We use M_i and Y_i to denote the observed value of the mediator and the outcome of interest for unit i, respectively. Under the formal statistical framework of causal inference (Neyman, 1923; Rubin, 1974; Holland, 1986), we write $M_i(t)$ to represent the potential mediator value under the treatment status $t = 0, 1$ where the observed value M_i equals the potential value of the mediator under the observed treatment status $M_i(T_i)$. Similarly, we use $Y_i(t, m)$ to denote the potential outcome under the treatment status t and the mediator value m where the observed outcome Y_i equals $Y_i(T_i, M_i(T_i))$.

In the literature of causal mediation analysis initiated by Robins and Greenland (1992) and Pearl (2001), the causal mediation effect (or indirect effect) for unit i given the treatment status t is defined as,

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0)), \tag{1}$$

which represents the causal effect of the treatment on the outcome that can be attributed to the

treatment-induced change in the mediator. This quantity represents the change in outcome under the scenario where the treatment variable is held constant at t and the mediator is changed from $M_i(0)$ to $M_i(1)$. Similarly, the unit-level direct effect of the treatment is defined as,

$$\zeta_i(t) \equiv Y_i(1, M_i(t)) - Y_i(0, M_i(t)), \tag{2}$$

which denotes the causal effect of the treatment on the outcome that can be attributed to the causal mechanisms other than the one represented by the mediator. Here, the mediator is held constant at $M_i(t)$ and the treatment variable is changed from 0 to 1.

Finally, the sum of these direct and indirect effects equals the following total effect, which formally decomposes the total effect into the direct and indirect effects,

$$\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \delta_i(t) + \zeta_i(1-t). \tag{3}$$

If we assume that the indirect and direct effects do not depend on the treatment status, i.e.,

$$\delta_i(t) = \delta_i \quad \text{and} \quad \zeta_i(t) = \zeta_i, \tag{4}$$

for each t, then we have simpler decomposition $\tau_i = \delta_i + \zeta_i$. Given these unit-level causal quantities of interest, we can define the population average effect for each quantity,

$$\bar{\delta}(t) = \mathbb{E}(\delta_i(t)), \quad \bar{\zeta} = \mathbb{E}(\zeta_i(t)), \quad \text{and} \quad \bar{\tau} = \mathbb{E}(\tau_i). \tag{5}$$

The goal of causal mediation analysis is, therefore, to decompose the total treatment effect into the direct and indirect effects where the former corresponds to the causal mechanism of interest and the latter represents all other causal mechanisms.

The standard mediation analysis most commonly used across various social science disciplines entails the process of fitting the following two linear regressions separately,

$$M_i = \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \varepsilon_{i2}, \tag{6}$$

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \varepsilon_{i3}, \tag{7}$$

where X_i represents a vector of observed *pre-treatment* confounders. After fitting these two models, researchers compute the product of two coefficients, i.e., $\hat{\beta}_2\hat{\gamma}$, and interpret it as an estimate of the average causal mediation effect (ACME) $\bar{\delta}(t)$ whereas the estimated coefficient $\hat{\beta}_3$ is interpreted as an

estimate of the average direct effect (ADE) $\bar{\zeta}(t)$. Alternatively, researchers can fit the regression model given in equation (7) as well as the following regression model,

$$Y_i = \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \varepsilon_{i1} \tag{8}$$

and compute the difference of the two coefficients, i.e., $\hat{\beta}_1 - \hat{\beta}_3$, to obtain the estimated average mediation effect. Here, $\hat{\beta}_1$ is taken as an estimate of the average total effect.

Imai et al. (2010c) prove that this standard mediation analysis can be justified under the following sequential ignorability assumption,

$$\{Y_i(t, m), M_i(t')\} \perp\!\!\!\perp T_i \mid X_i = x \tag{9}$$

$$Y_i(t', m) \perp\!\!\!\perp M_i \mid T_i = t, X_i = x \tag{10}$$

for any value of x, t, t', m and every unit i. In fact, it has been shown that under this assumption the ACMEs are nonparametrically identified, i.e., without any functional form or distributional assumptions. Imai et al. (2010a) develop general algorithms to compute an estimate of the ACME and its uncertainty given any statistical models specified by applied researchers (linear, nonlinear, parametric, nonparametric models, etc.).

3.2 Assumed Independence of Multiple Causal Mechanisms

What does the above assumption of sequential ignorability imply? At a first glance, the assumption may look identical to a set of the standard exogeneity assumptions. In fact, equation (9) implies the exogenous treatment assignment where the treatment is assumed to be randomized conditional on a vector of the observed pre-treatment variables X_i. However, the assumption about the mediator is different from the standard exogeneity assumption. According to equation (10), the exogeneity of the mediator must hold only after conditioning on the treatment and the observed pre-treatment variables. This means that even if the mediator can be assumed to be exogenous after conditioning on a vector of observed post-treatment confounders (denoted by W_i) as well as T_i and X_i, sequential ignorability is violated and the standard causal mediation analysis cannot identify the average mediation effects. Since alternative mediators are by definition causally affected by the treatment and influence the outcome, the assumption of no post-treatment confounding is equivalent to assuming that these alternative mediators do not causally influence the mediator of interest.

Going back to Figure 1, this means that the standard analysis based on the sequential ignorability assumption implicitly presupposes a situation like Figure 1(a), where the causal relationship between

the mediator of interest M_i and alternative mediators W_i is absent. In contrast, Figure 1(b) corresponds to a case where alternative mediators W_i causally affect the mediator of interest M_i. It is easy to see that the latter scenario violates the sequential ignorability assumption because W_i is a set of post-treatment variables that confound the relationship between M_i and Y_i. Thus, the standard causal mediation analysis assumes the independence between the causal mechanism of interest and other alternative mechanisms. As shown below, another difficulty resulting from this limitation of the standard causal mediation analysis is that this independence assumption is not directly testable from the observed data.

We now formalize the above argument. To make the existence of alternative mediators explicit, we introduce potential values of those other mediators W_i and denote them as $W_i(t)$ for $t = 0, 1$, implying that W_i is a post-treatment variable and hence possibly affected by the treatment. Since there is no causal relationship between M_i and W_i, the potential value of M_i is not a function of W_i and thus can be written as before, i.e., $M_i(t)$. The potential outcomes, on the other hand, depend on both M_i and W_i and are denoted by $Y_i(t, m, w)$ for any t, m, w. The observed outcome Y_i equals $Y_i(T_i, M_i(T_i), W_i(T_i))$. We emphasize that W_i could be either a single alternative mediator or a vector of multiple alternative mediators, and that those mediators may or may not be causally related to each other. Our framework does not require that researchers specify causal relationships among these alternative mechanisms.

Under this setup, we can define the two types of causal mediation effects, one with respect to M_i and the other with respect to W_i,

$$\delta_i^M(t) \equiv Y_i(t, M_i(1), W_i(t)) - Y_i(t, M_i(0), W_i(t)), \tag{11}$$

$$\delta_i^W(t) \equiv Y_i(t, M_i(t), W_i(1)) - Y_i(t, M_i(t), W_i(0)), \tag{12}$$

for $t = 0, 1$ where $\delta_i^M(t)$ ($\delta_i^W(t)$) represents the unit-level indirect effect of the treatment on the outcome through the mediator M_i (W_i) while holding the treatment at t and the other mediator at its value that would be realized under the same treatment status, i.e., $W_i(t)$ ($M_i(t)$). For example, in framing experiments such as Druckman and Nelson (2003) and Slothuus (2008), $\delta_i^M(t)$ represents the effect of issue frames on opinions that goes through changes in the perceived importance of the issue induced by the framing effect, while $\delta_i^W(t)$ equals the portion of the framing effect that operates through changes in belief content. As before, for each of these two causal mediation effects, we can define the ACME as $\bar{\delta}^M(t) \equiv \mathbb{E}(\delta_i^M(t))$ or $\bar{\delta}^W(t) \equiv \mathbb{E}(\delta_i^W(t))$ by averaging it over the target population $\mathcal{P}$.

In addition, the unit-level direct effect can be defined as,

$$\zeta_i(t, t') \equiv Y_i(1, M_i(t), W_i(t')) - Y_i(0, M_i(t), W_i(t')), \tag{13}$$

for each $t, t' = 0, 1$, where $\zeta_i(1,0)$, for example, represents the direct effect of the treatment while holding the mediators at $(M_i(1), W_i(0))$. Again, in Druckman and Nelson's experiment, this quantity represents the effect of frames that does not operate either through the importance mechanism or the content mechanism. As before, the ADEs are given by $\bar{\zeta}(t,t') \equiv \mathbb{E}(\zeta_i(t,t'))$. Given these definitions, we can decompose the total effect as the sum of direct and mediation (indirect) effects,

$$\begin{aligned} \tau_i &\equiv Y_i(1, M_i(1), W_i(1)) - Y_i(0, M_i(0), W_i(0)) && (14) \\ &= \delta_i^M(t) + \delta_i^W(1-t) + \zeta_i(1-t, t) && (15) \end{aligned}$$

As in the case of a single mediator, we can simplify this expression under the no-interaction assumptions,

$$\delta_i^M(t) = \delta_i^M, \quad \delta_i^W(t) = \delta_i^W, \quad \text{and} \quad \zeta_i(t,t') = \zeta_i, \tag{16}$$

for any t, t'. Under these conditions, we have a simpler decomposition relationship, i.e., $\tau_i = \delta_i^M + \delta_i^W + \zeta_i$.

We now generalize the sequential ignorability assumption to the case with multiple mediators where the causal independence between the main mediator of interest M_i and alternative mediators W_i is assumed.

Assumption 1 (Sequential Ignorability with Multiple Causally Independent Mediators) *We assume that the following three conditional independence statements hold,*

$$\begin{aligned} \{Y_i(t,m,w), M_i(t'), W_i(t'')\} &\perp\!\!\!\perp T_i \mid X_i = x, && (17) \\ Y_i(t', m, W_i(t')) &\perp\!\!\!\perp M_i \mid T_i = t, X_i = x, && (18) \\ Y_i(t', M_i(t'), w) &\perp\!\!\!\perp W_i \mid T_i = t, X_i = x, && (19) \end{aligned}$$

where $0 < \Pr(T_i = t \mid X_i = x)$ *and* $0 < p(M_i = m, W_i = w \mid T_i = t, X_i = x)$ *for any* x, t, t', m, w.

Under this assumption, it can be shown that the ACMEs, $\bar{\delta}^M(t)$ and $\bar{\delta}^W(t)$, as well as the ADEs $\bar{\zeta}(t,t')$ are nonparametrically identified and expressed by the same formula as Theorem 1 of Imai et al. (2010c). Proof of this identification result is provided in Appendix A.1.

Can the assumption of no causal relationship between these mediators be tested using the observed data? We note that Assumption 1 neither implies nor is implied by the conditional independence between the observed values of M_i and W_i given the treatment T_i and observed pre-treatment confounders X_i. This means that there exists no direct test of the assumed independence between causal mechanisms. However, we suggest that researchers at least check the degree of statistical dependence between M_i and W_i given (T_i, X_i) because the strong dependence between them is likely to indicate the violation of Assumption 1.

Next, we show that the standard analysis of multiple mediators in many social science applications, such as those described in Section 2, implicitly relies on Assumption 1. A common procedure to estimate mediation effects for multiple mechanisms is to fit the following set of linear regression models separately,

$$M_i = \alpha_M + \beta_M T_i + \xi_M^\top X_i + \varepsilon_{iM}, \tag{20}$$

$$W_i = \alpha_W + \beta_W T_i + \xi_W^\top X_i + \varepsilon_{iW}, \tag{21}$$

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \theta^\top W_i + \xi_3^\top X_i + \varepsilon_{i3}. \tag{22}$$

Then, the product of the estimated regression coefficients that correspond to the mechanism of interest (e.g. $\hat{\beta}_M \hat{\gamma}$) is computed and interpreted as an estimate of the ACME for that mechanism. The three studies of framing effects described in Section 2 all use this path-analytic approach and investigate the statistical significance of the estimated ACMEs for both mechanisms (or their standardized values).

It can be shown that, under Assumption 1, the product of the unstandardized path coefficients in equations (20) – (22) can be justified as an estimate of the corresponding ACME. The result is proved in Appendix A.2. However, it is important to note that Assumption 1 is merely an alternative representation of the same assumption, i.e., sequential ignorability given in equations (9) and (10). Thus, under Assumption 1, the same ACME can also be estimated by applying the standard causal mediation analysis to one of the mechanisms at a time. That is, under Assumption 1, the product of coefficients based on equations (6) and (7) also consistently estimates the ACME with respect to M_i, and the same analysis can be conducted for W_i using the linear structural equations model analogous to equations (6) and (7).

Earlier, we recommended that applied researchers should check the conditional independence of M and W given T and X. If these two mediators are dependent, then it is likely that Assumption 1 is violated. Indeed, if the separate application of the standard causal mediation analysis procedure to M and W gives the results that are different from those obtained by using the model specified in equations (20) – (22), then Assumption 1 is unlikely to hold because W is causally affecting M. A straightforward way to explore this possibility in the linear structural equations framework is to regress M on (W, T, X) and conduct an F-test with respect to W to see whether M and W are correlated even after conditioning on T and X. If one finds a statistically significant relationship, M is likely to be causally related to W, thereby violating the key identification assumption of the standard causal mediation analysis.

In sum, from a causal inference perspective, the standard path-analytic procedure for multiple me-

diators does not fundamentally add anything to the simpler one-mechanism-at-a-time procedure. In fact, they are based on the same exact assumption about the causal independence of multiple mediators. Although the statistical dependence between two variables does not necessarily imply their causal dependence, we suggest that researchers examine the statistical relationship between the mediator of interest and alternative mediators. Strong statistical dependence between those mediators implies that the sequential ignorability assumption is likely to be violated and researchers need a different statistical methodology in order to conduct causal mediation analysis in the presence of causally dependent multiple mediators. Before we turn to this issue in Section 5, we analyze the framing experiments under the assumption of independent causal mechanisms.

4 Empirical Analysis under the Independence Assumption

We reanalyze the data from the three experiments maintaining the assumption of independence between mechanisms using the framework of causal mediation analysis outlined by Imai et al. (2011). We also conduct a sensitivity analysis, which addresses the potential violation of the sequential ignorability assumption by explicitly calculating how much the estimate of the ACME could change if the assumption is violated to a specified degree. Unlike the one proposed later in this paper, however, the limitation of this existing sensitivity analysis is that it only allows for the possible existence of unobserved pre-treatment confounders and assumes no post-treatment confounder. In other words, the sensitivity analysis maintains the assumption of independent causal mechanisms and relaxes the exogeneity of mediator by introducing a certain degree of pre-treatment confounding. Below, we apply this procedure to the three framing experiments. We find that their original conclusions are largely (though less conclusively for some of the studies) valid and robust to unobserved confounders, *so long as we maintain the independence assumption*. Unfortunately, we also find evidence suggesting that alternative causal mechanisms are causally dependent on each other.

4.1 Druckman and Nelson (2003)

In their analysis, Druckman and Nelson find that the framing effect is mediated by the importance mechanism but not by the belief content mechanism. For the group of subjects who were not allowed to discuss the issue with other subjects, "the frame shaped the belief importance ratings, which in turn substantially affected overall opinions," whereas "the frames had minimal impact on the content measures and even when they did, [...] this effect did not carry through to overall opinions" (p.737). However, even under the assumption of independent mechanisms, a potential problem is that the estimated mediation effects may be biased due to unobserved pre-treatment confounding between the mediator and the outcome variable. For example, participants with libertarian ideology may think

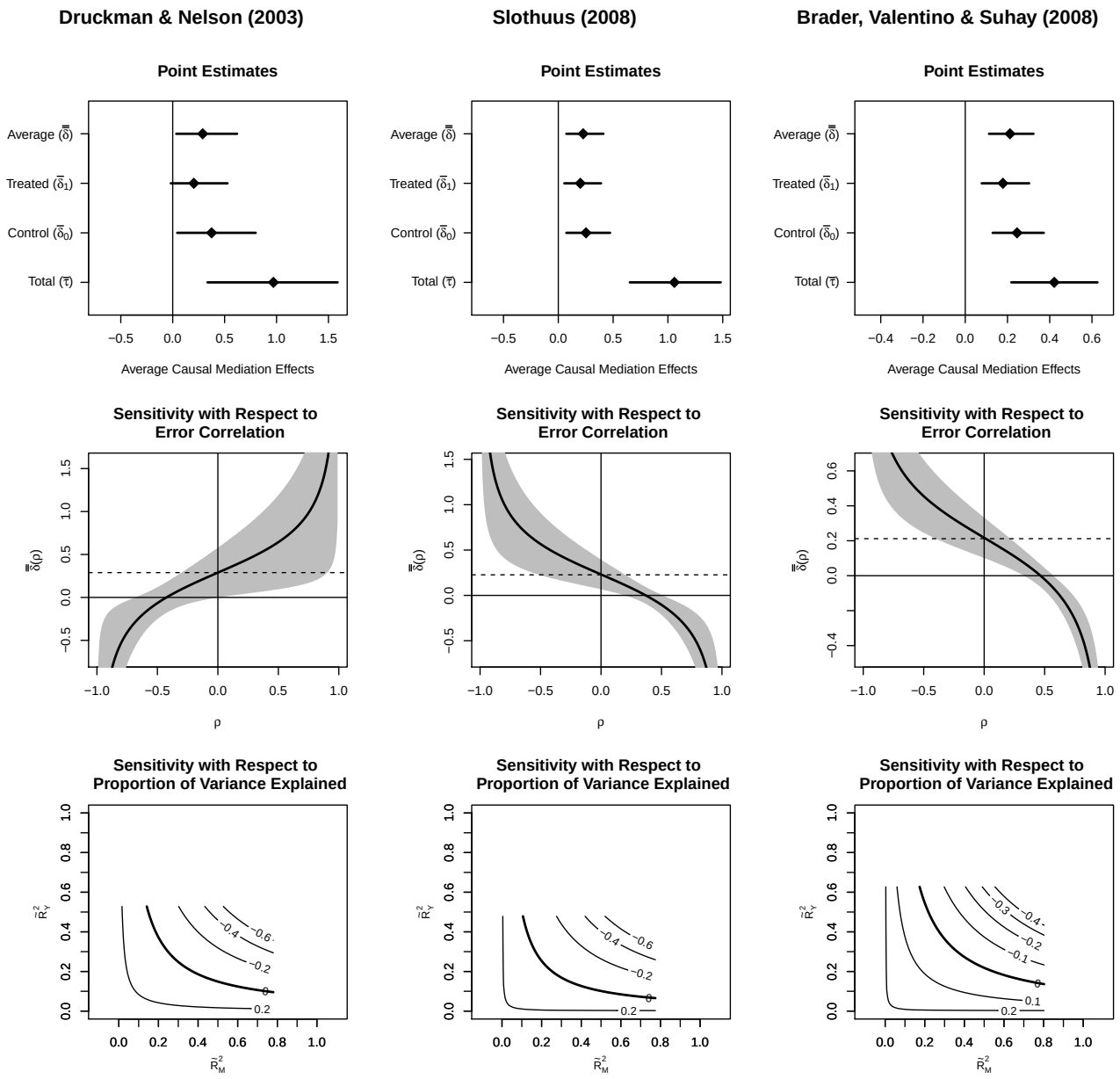


Figure 2: Estimation and Sensitivity Analysis Under the Assumption of Independent Causal Mechanisms. The top panels present the estimated ACMEs under the sequential ignorability assumption (Assumption 1) along with their 90% confidence intervals for each of the three studies indicated at the top. The panels in the middle row show the estimated true values of ACMEs as functions of the sensitivity parameter ρ, which represents the correlation between the error terms in the mediator and outcome models. The thick lines and gray bands represent the point estimates of the ACME and their 90% confidence intervals, respectively. The bottom panels show the same sensitivity analyses, except that the ACME estimates are plotted against $(\tilde{R}^2_Y, \tilde{R}^2_M)$, the proportions of the total variance in the outcome and mediator variables, respectively, that would be explained by a hypothetical unobserved pre-treatment confounder. Overall, the results suggest that *under the assumption of independence between mechanisms*, the causal mediation effects are positive and moderately statistically significant in all of the three studies, and the estimates are fairly robust to the possible unobserved pre-treatment mediator-outcome confounding to varying degrees.

freedom of speech more important than non-libertarian participants, and as a result they may also be more opposed to the campaign finance reform. If the available pre-treatment covariates fail to adjust for libertarianism, the estimated causal mediation effect may be upward biased.

Estimation of Causal Mediation Effects. To address this concern about confounding, we apply the framework of Imai et al. (2011) and estimate ACME and examine its robustness to the violation of sequential ignorability due to an unobserved pre-treatment confounder of the mediator and outcome.[3] Here, we focus on the no-discussion group and the causal mechanism corresponding to the perceived importance of freedom of speech (hereafter the "free speech importance" mechanism), which the original study found the most significant. As shown above, ignoring the other mechanisms should not affect the result under Assumption 1. The results are presented in the first column of Figure 2. The top figure presents the estimated ACME under sequential ignorability along with its 90% confidence interval. In addition to the overall estimate shown at the top, we also estimate the ACMEs for the treatment and control conditions separately in the plot in order to allow for the possibility that the ACME may differ depending on the baseline treatment status.

We find that the frame difference affected the participants' support for the campaign finance reform by approximately 0.286 points on the 7-point scale via the free speech importance mechanism, with the 90% confidence interval of $[0.025, 0.625]$. Because the average total causal effect of the frame difference is estimated to be 0.969 points, it suggests that about 28.6 percent of the total effect was transmitted through changes in the perceived importance of free speech. The ACME, however, appears to slightly differ depending on the baseline value of the treatment, as the estimate for the treatment condition (0.197) is closer to zero than the estimate under the control condition (0.375). In fact, the 90% confidence interval for the former overlaps with zero ($[-0.019, 0.497]$) while the interval for the latter does not ($[0.035, 0.819]$). Overall, though, the result confirms the original finding that the free speech importance mechanism significantly mediates the framing effect in the theoretically expected direction, although the estimation uncertainty is relatively large.

Sensitivity Analysis. The remaining two plots show the results of the sensitivity analysis with respect to two alternative sensitivity parameters. First, we calculate the estimated ACME as a function of the ρ parameter, the correlation between the error terms in the mediator and outcome models. A large value of ρ indicates existence of strong confounding between the mediator and outcome and thus a serious violation of the sequential ignorability assumption. The result suggests that the ACME equals

[3] We also include a set of observed pre-treatment covariates in the model in order to make the sequential ignorability assumption as plausible as the data permit. These covariates are year in college, age, gender, ethnicity, level of political knowledge, tendency for on-line processing, partisanship, and ideology.

zero when ρ is below -0.43, indicating a moderate degree of robustness compared to other similar studies (see Imai et al., 2011). The lower bound of the 90% confidence band, however, immediately crosses the zero line once we allow a slight negative correlation between the errors. Thus, a larger sample size may be needed in order to establish the robustness of the original findings to the possible existence of an unobserved pre-treatment confounder that affects the mediator and outcome in different directions (e.g., libertarianism).

Finally, the bottom plot shows the estimated true ACME as contour lines with respect to the $\tilde{R}^2_M$ and $\tilde{R}^2_Y$ parameters, the proportions of the total variance in the mediator and outcome variables, respectively, that would be explained by an unobserved pre-treatment confounder. The contours correspond to the scenario that the unobserved confounder affects the mediator and outcome in opposite directions (i.e., ρ is negative), as it is the only case where the estimated ACME can become negative. The result shows that the ACME can be estimated negative if the product of the two parameters are greater than 0.078. For example, the estimated ACME will be exactly zero if the unmeasured libertarianism explains 37% of the variation in the perceived importance of freedom of speech and 21% of the variation in the campaign finance reform opinions.

Discussion. The above results indicate that Druckman and Nelson's original conclusion about the free speech importance mechanism is largely valid under the assumption of independent causal mechanisms. The perceived importance of freedom of speech seems to mediate the framing effect, and the estimate is reasonably robust to the violation of the sequential ignorability assumption once we adjust for a set of pre-treatment covariates including partisanship and standard left-right ideology.

However, we emphasize that these results were obtained under the assumption that the free speech importance mechanism operates independently of other mechanisms, including the one represented by the participants' belief about the impact of the campaign finance reform on freedom of speech (hereafter the "free speech belief" mechanism). The assumption would be violated if, for example, the change in the content of participants' belief about the reform due to framing differences then caused any changes in their perceived importance of free speech. Indeed, this is of major concern because, as Miller (2007) points out on the basis of her experimental study, "individuals use information obtained from the media to evaluate how important issues are" (p.711) and "when media exposure to an issue causes negative emotional reactions about the issue, increased importance judgments will follow" (p.712).

The data in fact underline the concern. By regressing the free speech importance mediator on the belief content mediator, the treatment, and the other covariates, we find that the coefficient of the belief mediator is negative and significantly different from zero at 0.1 level (-0.23, with the p-value of 0.093). As discussed earlier, this analysis suggests that there may exist an additional causal arrow linking the

free speech belief mechanism to the importance mechanism. Such a dependent mediator can be seen as a *post-treatment confounder*, whose existence, whether observed or unobserved, can cause a substantial bias in the estimate of ACME and invalidate the above results (see also Appendix of Imai et al., 2011).

4.2 Slothuus (2008)

Contrary to Druckman and Nelson (2003), Slothuus (2008) finds both importance and content mechanisms to be at work and concludes "both the importance change process and the content change process mediate the framing effects" (p.18). Like Druckman and Nelson, Slothuus relies on a path-analytic method to estimate the mediation effects. We therefore estimate ACME for the importance mechanism and evaluate its robustness to unobserved pre-treatment confounding between the mediator and outcome variable under the assumption of independent causal mechanisms.[4] For simplicity, we focus on one of the mediators used in the original analysis (the "incentive to work" importance) which was found to be the most statistically significant.

Estimation of Causal Mediation Effects. The results are shown in the middle column of Figure 2. The top panel shows that the ACME for the incentive-to-work importance mechanism is estimated to be about 0.230 on the 7-point scale, with the 90% confidence interval of $[0.082, 0.402]$. This represents approximately 21.3% of the total framing effect, which is estimated to be 1.064 points ($[0.640, 1.49]$). Unlike the Druckman and Nelson study, this proportion does not appear to vary depending on the baseline treatment frame: the estimated ACME is largely similar for the treatment (0.205, $[0.052, 0.400]$) and control (0.255, $[0.092, 0.445]$) conditions. Overall, the result confirms the original finding that the importance mechanism significantly mediates the framing effect for the social welfare reform.

Sensitivity Analysis. How robust is this conclusion to the possible existence of unobserved pre-treatment confounding? The middle panel shows that the ACME for the importance mechanism equals zero when the error correlation between the mediator and outcome models becomes greater than 0.37. This indicates a degree of robustness largely comparable to that of the Druckman and Nelson study. However, the estimated ACME has smaller estimation uncertainty and hence the lower confidence bound does not cross zero until the ρ parameter becomes greater than 0.20. The analysis with respect to the $\tilde{R}^2$ parameters leads to similar conclusions, as shown in the bottom panel. The estimated ACME becomes negative when the product of $\tilde{R}^2_M$ and $\tilde{R}^2_Y$ is greater than 0.05, which is again comparable to the result of the Druckman and Nelson study. For example, the estimate equals zero if an unobserved pre-treatment confounder explains 25% of the variation in the perceived importance of work incentives

[4] Again, we add a set of observed pre-treatment covariates to the model to make the sequential ignorability assumption as plausible as possible. The covariates are gender, education, level of political interest, self-placement on a left-right ideology scale, school, year of birth, political knowledge, and extremity of political values.

and 20% of the variation in the welfare reform opinion.

Discussion. The Slothuus study is subject to the same potential problem as the Druckman and Nelson study because both are based on the assumption that alternative causal mechanisms are independent of one another. For example, the change in the content of participants' considerations about the welfare reform may not only directly affect their opinions but also influence the perceived importance of work incentives, thereby also affecting the outcome indirectly through further changes in the importance mediator. If this were to be the case, there exists a causal arrow linking the content mediator to the importance mediator, violating the independence assumption. Indeed, regressing the importance mediator on the content mediator and the treatment (as well as the set of pre-treatment covariates), we find a large positive and statistically significant coefficient on the content mediator (0.350, with the p-value of 0.000). Thus, similar to the Druckman and Nelson study, the possibility of bias due to post-treatment confounding is also an important concern.

4.3 Brader, Valentino and Suhay (2008)

The original analysis of Brader et al. (2008) reveals that the combination of Latino cues and negative framing changed the participants' opinions and behavior in the anti-immigration direction. That is, the participants who read the Latino article emphasizing the negative consequences of immigration became more opposed to increased immigration than the rest of the participants. More importantly, the authors find that the frame affected the outcome variables through the anxiety mechanism rather than the perceived harm mechanism.[5] Like the other two studies, Brader et al. rely on the structural equation modeling approach of Baron and Kenny (1986). However, the validity of this causal mediation analysis crucially hinges on the sequential ignorability assumption. The estimated effects will be biased if, for example, unmeasured job skills of participants affect both their levels of anxiety and opinions about increased immigration. This is likely if low-skill workers feel more anxious about immigrants and are also more anti-immigration than high-skill participants.

Estimation of Causal Mediation Effects. We estimate the ACME for the anxiety mechanism by focusing on one of the outcome variables in the original study (whether immigration to the United States should be increased or decreased) and examine how much of the total effect of the negative Latino frame on the variable operated through the anxiety mediator. The point estimates are shown in the top panel of the left column in Figure 2 along with their 90% confidence intervals. The result supports the original conclusion, indicating that the negative Latino frame made the participants 0.216

[5] In their own words, "The conjunction of Latino cues and negative news about immigration influenced levels of anxiety. Anxiety then caused shifts in policy attitudes, information seeking, and political action" (p.969). In contrast, "belief about the severity of the immigration problem does not mediate the interactive effect of ethnic cues and news emphasis" (p.969).

points more opposed to immigration on the 5-point scale via the anxiety mechanism, with the 90% confidence interval of ($[0.120, 0.329]$). This represents roughly half of the estimated total framing effect, which is 0.423 points. The estimate varies slightly across the treatment values (0.184 with $[0.083, 0.311]$ for the treated; 0.247 with $[0.136, 0.360]$ for the control).

Sensitivity Analysis. The middle and bottom plots show the result of sensitivity analysis. According to this analysis, the ACME is estimated to be negative when the correlation between the error terms in the mediator and outcome models is larger than 0.47, and the ACME is statistically indistinguishable from zero at the 90% level when the ρ parameter is greater than 0.33. Thus, the result of this study is slightly more robust to the sequential ignorability violation than those of the two above studies. Finally, in terms of the coefficient of determination parameters, the product of $\tilde{R}^2_M$ and $\tilde{R}^2_Y$ must be at least as large as 0.08, implying that an unobserved pre-treatment confounder must explain 20% of the variation in the participants' anxiety and 40% of the variation in their immigration opinions, for example. These results confirm the original finding that the anxiety mechanism plays a major role in the framing effect on opinions toward increased immigration.

Discussion. As in the previous two studies, the above analysis rests on the key assumption that the anxiety mechanism is independent of other mechanisms underlying the framing effect. In particular, the participants' conscious belief about the negative impact of increased immigration — the other mechanism explicitly studied in the original article — is assumed to have no impact on their levels of anxiety. This assumption may not be entirely plausible because the increased level of perceived harm of immigration due to the negative Latino framing can cause the participants to feel more anxious about increase in immigration (Isbell and Ottati, 2002). Indeed, there is a strong statistical dependence between the two mediators. Regressing the anxiety mediator on the perceived harm mediator as well as the treatment and the pre-treatment covariates, we find that the estimated coefficient of the perceived harm mediator is positive and large (1.016, with the p-value of 0.000). Thus, like in the above two examples, we must address the issue of the post-treatment confounding between the mediator and outcome variable in order to evaluate the empirical findings of this study.

5 Statistical Analysis of Causally Related Multiple Mechanisms

In this section, we consider statistical analysis of causally related multiple mechanisms. In particular, we focus on the situation depicted in Figure 1(b) where other mediators W confound the relationship between the mediator of interest M and the outcome Y. We assume the exogeneity of the mediator M conditional on the pre-treatment covariates X, the treatment T, and the post-treatment confounders W. In this setting, we first review the identification result of Robins (2003) that the average causal

mediation effects (represented by red arrows in the figure) can be nonparametrically identified under an additional assumption of no interaction between the treatment T and the mediator M.

However, as noted by many researchers (e.g. Petersen et al., 2006; Imai et al., 2012), this assumption is unrealistic in many applications. In particular, the assumption must hold not just on average but rather for every observation. To overcome this limitation, we propose a new sensitivity analysis, which can be used to examine how empirical results change as we gradually relax the no-interaction assumption. Such an analysis reveals how robust one's empirical results are to the potential violation of the key identification assumption. We develop this methodology in the context of a fairly general varying-coefficient linear regression model where each coefficient can vary arbitrarily across observations. We then apply the proposed method to the framing experiments introduced earlier.

5.1 The Setup and Assumptions

We use the framework of causal mediation analysis and its associated notation introduced in Sections 3.1 and 3.2. Recall that the key difference between Figures 1(a) and 1(b) is that we allow the mediator of interest M to be causally affected by a set of alternative mediators W. This means that the potential values of M are now a function of W. That is, we use $M_i(t, w)$ to denote the potential value of the mediator of interest for unit i when the treatment status is t and the alternative mediators W take the value of w. Then, the observed value of the mediator for this unit is given by $M_i = M_i(T_i, W_i(T_i))$.

Under this setting, for each unit, we define the causal mediation (with respect to M) and direct effects as,

$$\delta_i(t) \equiv Y_i(t, M_i(1, W_i(1)), W_i(t)) - Y_i(t, M_i(0, W_i(0)), W_i(t)) \tag{23}$$

$$\zeta_i(t) \equiv Y_i(1, M_i(t, W_i(t)), W_i(1)) - Y_i(0, M_i(t, W_i(t)), W_i(0)) \tag{24}$$

for $t = 0, 1$ where $\delta_i(t)$ corresponds to the causal effect of the treatment on the outcome that transmits through the mediator of interest M (i.e., the red arrows in Figure 1(b)). In the framing experiment of Druckman and Nelson, for example, this represents the portion of the framing effect due to the change in issue importance induced by the framing manipulation, while the belief content is held constant at the value that would be naturally observed when one of the issue frame is given. On the other hand, $\zeta_i(t)$ represents the rest of the treatment effect (denoted by the black arrow at the bottom of the figure and the combination of the red and black arrows that go from T to Y through W but not through M). In Druckman and Nelson's experiment, this represents the fraction of the framing effect that does not go through the issue importance mechanism, regardless of whether it gets transmitted through the belief content mechanism or through other unspecified mechanisms. Notice that these two

effects represent the quantities identical to those in equations (1) and (2); the only difference is that the expressions in equations (23) and (24) make the existence of the alternative mediators W explicit while equations (1) and (2) do not. Thus, as expected, the sum of these two effects equals the total treatment effect,

$$\tau_i \equiv Y_i(1, M_i(1, W_i(1)), W_i(1)) - Y_i(0, M_i(0, W_i(0)), W_i(0)) = \delta_i(t) + \zeta_i(1-t) \tag{25}$$

for $t = 0, 1$. Again, we are interested in estimating the ACME, i.e., $\bar{\delta}(t) \equiv \mathbb{E}(\delta_i(t))$, and the average direct and total effects can also be defined in an analogous manner, i.e., $\bar{\zeta}(t) \equiv \mathbb{E}(\zeta_i(t))$ and $\bar{\tau} \equiv \mathbb{E}(\tau_i)$.

What assumptions do we need to make in order to identify the ACME in this scenario? We modify Assumption 1 and consider the following weaker version of the sequential ignorability assumption,

Assumption 2 (Sequential Ignorability with Multiple Causally Dependent Mediators) *We assume that the following three conditional independence statements hold,*

$$\{Y_i(t, m, w), M_i(t, w), W_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x \tag{26}$$

$$\{Y_i(t, m, w), M_i(t, w)\} \perp\!\!\!\perp W_i \mid T_i = t,\ X_i = x \tag{27}$$

$$\{Y_i(t, m, w)\} \perp\!\!\!\perp M_i \mid W_i(t) = w,\ T_i = t,\ X_i = x \tag{28}$$

for any t, m, w, x.

This assumption corresponds to what Robins (1986, 2003) called the FRCISTG (fully randomized causally interpretable structural tree graph) model. Assumption 2 is similar to Assumption 1 in that exogeneity is assumed for the treatment T, the alternative mediators W, and the mediator of interest M.

However, Assumption 2 relaxes Assumption 1 in two important ways. First, the mediator of interest M is assumed to be exogenous after conditioning on alternative mediators W as well as the treatment T and the pre-treatment confounders X whereas Assumption 1 does not permit conditioning on the post-treatment confounders. In the context of Druckman and Nelson's framing experiment, this implies that Assumption 2 allows for the possibility that the perceived issue importance is affected by the content of the belief about the issue whereas Assumption 1 rules out the existence of such causal dependence.

Second, Assumption 2 avoids specifying the conditional independence relationship between the potential outcome under the treatment status t and the potential value of mediator under the opposite treatment status t'. For example, contrast equation (18) with equation (28). The former assumes the conditional independence between $Y_i(t', m, W_i(t'))$ and $M_i(t)$ even when $t \neq t'$ whereas the latter only applies the conditional independence assumption to the relationship between $Y_i(t, m, w)$ and $M_i(t)$.

Some scholars (e.g. Robins and Richardson, 2010) consider this distinction to be important because one can conceive of a series of randomized experiments that by design satisfy Assumption 2, while such experiments do not exist for Assumption 1 (see Section 7) even as a purely theoretical possibility. This implies that Assumption 2 can be empirically verified from observed data at least in theory, while Assumption 1 cannot.

Unfortunately, in the presence of causally dependent multiple mediators, Assumption 2 is not sufficient for the identification of the ACME. Robins (2003) shows that under this setting the ACMEs are nonparametrically identifiable if the following assumption of no treatment-mediator interaction effect holds:

Assumption 3 (No Interaction Between Treatment and Mediator) *For every unit i, we assume the following equality,*

$$Y_i(1, m, W_i(1)) - Y_i(0, m, W_i(0)) \quad = \quad Y_i(1, m', W_i(1)) - Y_i(0, m', W_i(0)), \tag{29}$$

for any m, m'.

The problem of this assumption is that it is unlikely to be credible in most applications because it must hold for every unit. In Druckman and Nelson's experiment, for example, Assumption 3 implies that for every subject, the causal effect of the perceived issue importance on opinions must be constant regardless of whether the subject read the positive or negative article. To overcome this limitation, we now introduce our proposed methodology that allows one to relax this no interaction assumption.

5.2 The Proposed Methodology

Given the above setup, we show how to relax the no-interaction assumption (Assumption 3) while maintaining the exogeneity of treatment and mediator (Assumption 2). We consider the following linear structural equation model with varying coefficients,

$$M_i(t, w) \quad = \quad \alpha_2 + \beta_{2i} t + \xi_{2i}^\top w + \mu_{2i}^\top tw + \lambda_{2i}^\top x + \varepsilon_{2i}, \tag{30}$$

$$Y_i(t, m, w) \quad = \quad \alpha_3 + \beta_{3i} t + \gamma_i m + \kappa_i tm + \xi_{3i}^\top w + \mu_{3i}^\top tw + \lambda_{3i}^\top x + \varepsilon_{3i}, \tag{31}$$

where $\mathbb{E}(\varepsilon_{2i}) = \mathbb{E}(\varepsilon_{3i}) = 0$ is assumed without the loss of generality.[6] The model generalizes the linear structural equation model commonly used by applied researchers (see equations (20) and (22)) in several important ways. First, the model reflects the particular situation depicted in Figure 1(b) and discussed above which permits the presence of causally dependent multiple mediators. Specifically, the mediator of interest M is allowed to depend on a set of alternative mediators W, which themselves are

[6] This model encompasses the model considered by Glynn (2012) as a special case.

possibly affected by the treatment, as well as the treatment variable T. Similarly, the outcome variable Y depends on W as well as M and T. Second, each coefficient is allowed to vary across individual observations in an arbitrary manner, allowing for a wide range of patterns of heterogeneous treatment effects. This is a crucial advantage over the traditional structural equation modeling framework, which typically assumes the unit homogeneity of treatment effects. Finally, we include the interaction between the treatment and each of the mediators (both M and W) so that mediation effects can vary depending on the baseline treatment status.[7]

How is this model related to the standard linear structural equation model? We decompose each of the varying coefficients into the mean and the deviation from it,

$$M_i(t,w) = \alpha_2 + \beta_2 t + \xi_2^\top w + \mu_2^\top tw + \lambda_2^\top x + \eta_{2i}(t,w), \tag{32}$$

$$Y_i(t,m,w) = \alpha_3 + \beta_3 t + \gamma m + \kappa tm + \xi_3^\top w + \mu_3^\top tw + \lambda_3^\top x + \eta_{3i}(t,m,w), \tag{33}$$

where $\beta_2 \equiv \mathbb{E}(\beta_{2i}), \beta_3 \equiv \mathbb{E}(\beta_{3i}), \gamma \equiv \mathbb{E}(\gamma_i), \kappa \equiv \mathbb{E}(\kappa_i), \xi_2 \equiv \mathbb{E}(\xi_{2i}), \xi_3 \equiv \mathbb{E}(\xi_{3i}), \mu_2 \equiv \mathbb{E}(\mu_{2i}), \mu_3 \equiv \mathbb{E}(\mu_{3i}), \lambda_2 \equiv \mathbb{E}(\lambda_{2i})$ and $\lambda_3 \equiv \mathbb{E}(\lambda_{3i})$ are the mean parameters of corresponding varying coefficients. The new error terms are given by,

$$\eta_{2i}(t,w) = \tilde{\beta}_{2i} t + \tilde{\xi}_{2i}^\top w + \tilde{\mu}_{2i}^\top tw + \tilde{\lambda}_{2i}^\top x + \varepsilon_{2i} \tag{34}$$

$$\eta_{3i}(t,m,w) = \tilde{\beta}_{3i} t + \tilde{\gamma}_i m + \tilde{\kappa}_i tm + \tilde{\xi}_{3i}^\top w + \tilde{\mu}_{3i}^\top tw + \tilde{\lambda}_{3i}^\top x + \varepsilon_{3i} \tag{35}$$

where by construction we have $\mathbb{E}(\tilde{\beta}_{2i}) = \mathbb{E}(\tilde{\beta}_{3i}) = \mathbb{E}(\tilde{\xi}_{2i}) = \mathbb{E}(\tilde{\xi}_{3i}) = \mathbb{E}(\tilde{\mu}_{2i}) = \mathbb{E}(\tilde{\mu}_{3i}) = \mathbb{E}(\tilde{\lambda}_{2i}) = \mathbb{E}(\tilde{\lambda}_{3i}) = \mathbb{E}(\tilde{\gamma}_i) = \mathbb{E}(\tilde{\kappa}_i) = 0$ and hence $\mathbb{E}(\eta_{2i}(t,w)) = \mathbb{E}(\eta_{3i}(t,m,w)) = 0$. Since Assumption 2 implies the following exogeneity conditions, $\mathbb{E}(\varepsilon_{2i} \mid X_i, T_i, W_i) = \mathbb{E}(\varepsilon_{3i} \mid T_i, W_i, M_i) = 0$, it follows that the exogeneity condition also holds for the new error terms, i.e., $\mathbb{E}(\eta_{2i}(T_i, W_i) \mid X_i, T_i, W_i) = \mathbb{E}(\eta_{3i}(T_i, M_i, W_i) \mid X_i, T_i, W_i, M_i) = 0$. Thus, the coefficients in equation (32) and (33) can be estimated without bias under Assumption 2.

We show that the ACMEs are identified if we fix two unobserved quantities, which we use as sensitivity parameters. The first parameter is the correlation between the mediator of interest $M_i(t)$ and the individual-level treatment-mediator interaction effect κ_i, i.e., $\rho_t = \mathrm{Corr}(M_i(t, W_i(t)), \kappa_i)$. This parameter represents the direction of the interaction effect. The second parameter is the standard deviation of the individual-level coefficient for the treatment-mediator interaction, i.e., $\sigma = \sqrt{\mathbb{V}(\kappa_i)}$, representing

[7] The proposed analysis can also be extended to the model including the interaction between the two mediators (M and W), though the resulting sensitivity analysis would include additional sensitivity parameters and thus become more complex. We thus focus on the simpler model assuming no interaction between M and W in equation (31).

the degree of heterogeneity in the treatment-mediator interaction effect. In Appendix A.3, we prove that the ACMEs and direct effects can be written as a function of the identifiable model parameters (under Assumption 2 but without requiring Assumption 3) and the two sensitivity parameters, for $t = 0, 1$,

$$\bar{\delta}(t) = \bar{\tau} - \bar{\zeta}(1-t) \tag{36}$$

$$\bar{\zeta}(t) = \beta_3 + \kappa \mathbb{E}(M_i \mid T_i = t) + \rho_t \sigma \sqrt{\mathbb{V}(M_i \mid T_i = t)} + (\xi_3 + \mu_3)^\top \mathbb{E}(W_i \mid T_i = 1) - \xi_3^\top \mathbb{E}(W_i \mid T_i = 0). \tag{37}$$

Thus, under the situation depicted in Figure 1(b) with Assumption 2, we can conduct a sensitivity analysis even in the presence of post-treatment confounders to examine how the estimated ACME changes as a function of the two parameters, ρ_t and σ. Several remarks are in order. First, the no-interaction assumption given in equation (29) implies that $\kappa_i = 0$ and hence $\kappa = \sigma = \rho_t = 0$. Thus, under the model considered here, the ACME and the average direct effect are identified as $\bar{\delta}(t) = \bar{\tau} - \beta_3 - (\xi_3 + \mu_3)^\top \mathbb{E}(W_i \mid T_i = 1) + \xi_3^\top \mathbb{E}(W_i \mid T_i = 0)$ and $\bar{\zeta}(t) = \beta_3 + (\xi_3 + \mu_3)^\top \mathbb{E}(W_i \mid T_i = 1) - \xi_3^\top \mathbb{E}(W_i \mid T_i = 0)$. These expressions correspond exactly to the procedure used by applied researchers who rely on a linear regression to model the conditional expectation of W_i given T_i (e.g., Taylor et al., 2008). While these researchers are unaware of the essential role of the no-interaction assumption, the sensitivity analysis developed here can formally assess the robustness of empirical results to the potential violation of this key identifying assumption.

Second, we can relax the no-interaction assumption of equation (29) to some extent under our linear structural equation framework by considering instead the following homogeneous interaction effect assumption,

$$Y_i(1, m, W_i(1)) - Y_i(0, m, W_i(0)) = B_i + Cm \tag{38}$$

for any m. Unlike the no-interaction assumption, this assumption allows for the treatment-mediator interaction in a way that is common to all observations. This homogenous interaction effect assumption implies $\sigma = 0$ and thus we can identify the ACME and average direct effect whose expressions are given by $\bar{\delta}(t) = \bar{\tau} - \bar{\zeta}(1-t)$ and $\bar{\zeta}(t) = \beta_3 + \kappa \mathbb{E}(M_i \mid T_i = t) + (\xi_3 + \mu_3)^\top \mathbb{E}(W_i \mid T_i = 1) - \xi_3^\top \mathbb{E}(W_i \mid T_i = 0)$, respectively.

Third, when neither of the above assumptions holds, the standard estimation procedure results in bias. For the ACME, this bias equals $-\rho_{1-t}\sigma\sqrt{\mathbb{V}(M_i \mid T_i = 1-t)}$, which depends on the variance of

mediator within the treatment or control group as well as the sensitivity parameters. For example, the bias is negative and large if the treatment-mediator interaction effect tends to be higher when the mediator takes a larger value, i.e., $\rho_t > 0$ and the variance of mediator and the degree of heterogeneity in these interaction effects are large.

Fourth, we note that when $\rho_t = 0$, we can identify the ACME regardless of the value of σ. However, when ρ_t is not equal to zero, we must specify both ρ_t and σ in order to estimate the ACME under Assumption 2. If the interpretation of ρ_t is difficult, we can derive the bounds on the ACME as a function of σ while allowing ρ_t to take any value between -1 and 1.[8]

Fifth, for the ease of the interpretation of σ, we follow Imai et al. (2010a,c) and use coefficients of determination as an alternative parameterization. Specifically, we use the proportion of the unexplained or original variance of the outcome that is explained by incorporating the heterogeneity in the treatment-mediator interaction. Thus, the sensitivity parameter represents how important it would be to incorporate the interaction heterogeneity in the regression model in order to explain the variation in the outcome variable. Formally, these parameters are defined as

$$R^{2*} \;=\; \frac{\mathbb{V}(\tilde{\kappa}_i T_i M_i)}{\mathbb{V}(\eta_{3i}(T_i, M_i, W_i))} \quad \text{and} \quad \widetilde{R}^2 \;=\; \frac{\mathbb{V}(\tilde{\kappa}_i T_i M_i)}{\mathbb{V}(Y_i)} \tag{39}$$

for the proportion of unexplained variance and that of the original variance explained by the heterogeneity of the treatment-mediator interaction effects, respectively. We can directly relate these quantities to the ACME via the following one-to-one relationship between σ and each of these coefficients of determination,[9]

$$\sigma \;=\; \sqrt{\frac{\mathbb{V}(\eta_{3i}(T_i, M_i, W_i))R^{2*}}{\mathbb{E}(T_i M_i^2)}} \;=\; \sqrt{\frac{\mathbb{V}(Y_i)\widetilde{R}^2}{\mathbb{E}(T_i M_i^2)}}. \tag{40}$$

This implies that σ is bounded from above by $\sqrt{\mathbb{V}(\eta_{3i}(T_i, M_i, W_i))/\mathbb{E}(T_i M_i^2)}$ because $0 \leq R^{2*} \leq 1$. Furthermore, the sensitivity to the interaction heterogeneity can be assessed by studying how the ACME varies depending on the values of R^{2*} and $\widetilde{R}^2$. This can also be done by calculating the ratio of σ to its upper bound.

Finally, under Assumption 2, we can also identify another possible quantity of interest, the popula-

[8] The expression for such bounds is given by the following, $\bar{\tau} - \beta_3 - \kappa\mathbb{E}(M_i \mid T_i = 1-t) - \sigma\sqrt{\mathbb{V}(M_i \mid T_i = 1-t)} - (\xi_3 + \mu_3)^\top \mathbb{E}(W_i \mid T_i = 1) + \xi_3^\top \mathbb{E}(W_i \mid T_i = 0) \;\leq\; \bar{\delta}(t) \;\leq\; \bar{\tau} - \beta_3 - \kappa\mathbb{E}(M_i \mid T_i = 1-t) + \sigma\sqrt{\mathbb{V}(M_i \mid T_i = 1-t)} - (\xi_3 + \mu_3)^\top \mathbb{E}(W_i \mid T_i = 1) + \xi_3^\top \mathbb{E}(W_i \mid T_i = 0)$.

[9] We used the following equality, $\mathbb{V}(\tilde{\kappa}_i T_i M_i) \;=\; \mathbb{E}(\mathbb{V}(\tilde{\kappa}_i T_i M_i \mid T_i, M_i, W_i)) + \mathbb{V}(\mathbb{E}(\tilde{\kappa}_i T_i M_i \mid T_i, M_i, W_i)) \;=\; \mathbb{E}(T_i M_i^2 \mathbb{V}(\tilde{\kappa}_i \mid T_i, M_i, W_i)) \;=\; \sigma^2 \mathbb{E}(T_i M_i^2)$ where the second equality is due to the law of total variance, and the next two equalities hold because of Assumption 2.

tion average of the causal mediation effect specific to the path $T \to W \to Y$. This quantity is defined as

$$\chi_i(t) \;=\; Y_i(t, M_i(t, W_i(t)), W_i(1)) - Y_i(t, M_i(t, W_i(t)), W_i(0)), \tag{41}$$

for $t = 0, 1$ and corresponds to the effect of the treatment on the outcome that goes through the post-treatment confounders W but not through the primary mediator M. In Appendix A.4, we prove that $\bar{\chi}(t) \equiv \mathbb{E}(\chi_i(t))$ is given by

$$\bar{\chi}(t) \;=\; (\xi_3 + t\mu_3)\left\{\mathbb{E}(W_i \mid T_i = 1) - \mathbb{E}(W_i \mid T_i = 0)\right\} \tag{42}$$

for $t = 0, 1$. It is noteworthy that under our model this identification result only requires Assumption 2 and does not need any assumption about the treatment-mediator interaction or sensitivity parameters. This contrasts with the result by Albert and Nelson (2011) that the path-specific effect $\bar{\chi}(t)$ cannot be identified under the set of assumptions they consider, as well as the more general nonparametric identification result about path-specific effects by Avin et al. (2005).

6 Empirical Analysis without the Independence Assumption

We now revisit the media framing experiments introduced in Section 2. The authors of these studies implicitly assumed that the mechanisms of their primary interest were causally independent from the alternative mechanisms incorporated in the original analyses. However, as discussed in Section 4, this assumption is implausible on both theoretical and empirical grounds. We therefore use the method developed in the previous section to estimate ACMEs without assuming the independence between these mechanisms. We also apply our proposed sensitivity analysis to assess how robust these estimates are to the violation of the key assumption that the effect of the primary mediator does not depend on the value of the treatment (Assumption 3).

6.1 Druckman and Nelson (2003)

In Section 4.1, we noted that in Druckman and Nelson's framing experiment, the participants' belief about the effect of the campaign finance reform may have influenced their perceived importance of free speech. We thus reanalyze their data to allow for the causal dependence of the importance mechanism on the belief content mechanism. That is, we first estimate the linear structural equations models in equations (32) and (33) with the same set of pretreatment covariates as in Section 4.1, and then compute the ACME as a function of the sensitivity parameters using equations (36) and (37). The

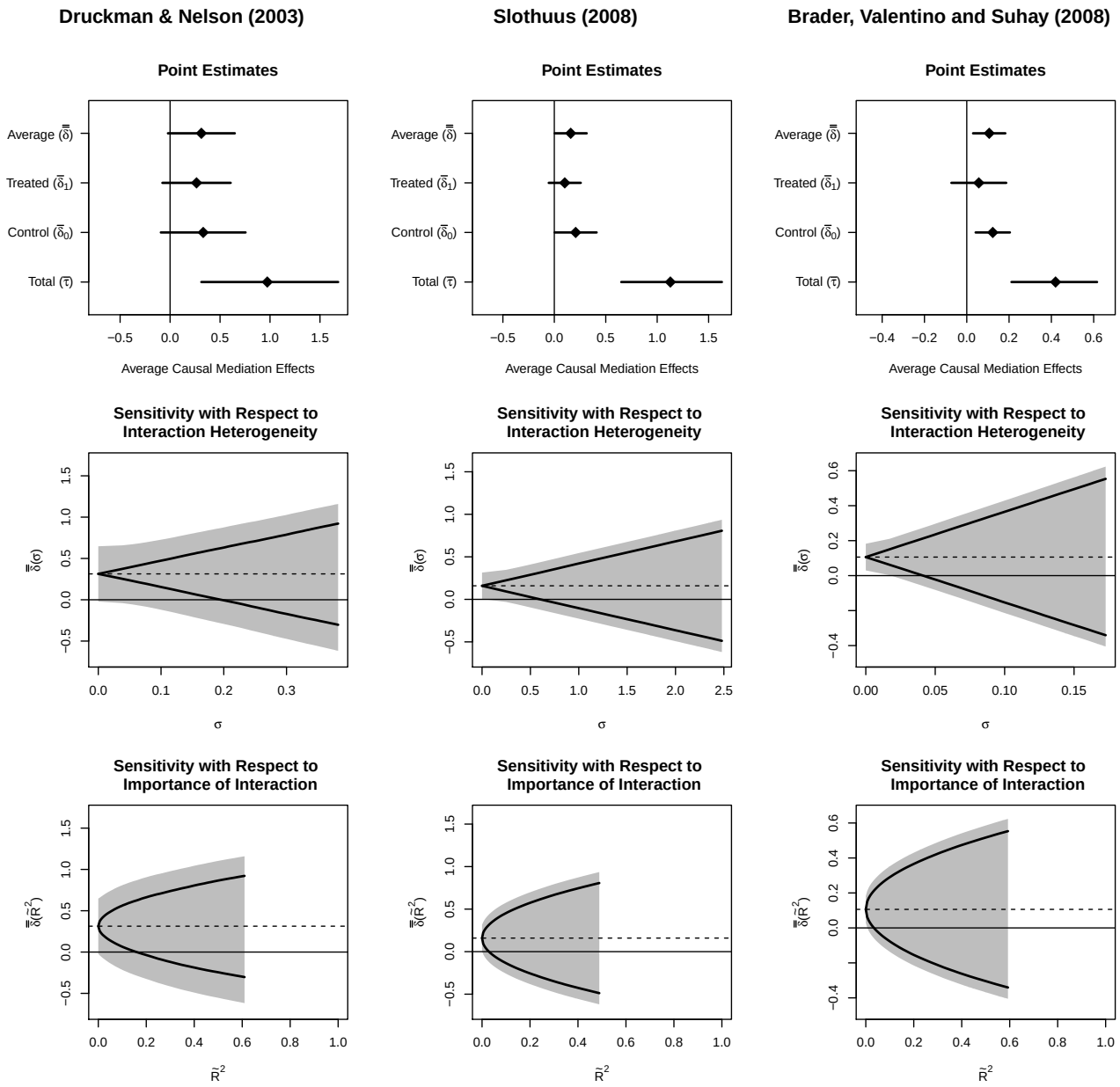


Figure 3: Estimation and Sensitivity Analysis Without the Independence Assumption. The top panels present the estimated ACMEs under the sequential ignorability and homogenous interaction assumptions (Assumption 2 and equation 38) with their 90% confidence intervals for the three framing studies. The panels in the middle row show the sharp bounds on the true values of ACMEs as functions of the sensitivity parameter σ, which equals the standard deviation of the varying coefficient on the treatment-mediator interaction term and thus represents the degree of unit heterogeneity in the interaction. The bottom panels plot the same estimated ACMEs with respect to $\widetilde{R}^2$, the proportion of the total variance of the outcome variable that would be explained by the treatment-mediator interaction term. The results suggest that the positive mediation effects found in the original studies (see Figure 2) become smaller in magnitude and statistically either insignificant or barely significant once we allow for the causal pathway from the alternative mediator (W) to the mediator of interest (M) as in Figure 1(b).

results are shown in the three plots in the left column of Figure 3.

As in Section 4, the top panel presents the estimated ACMEs (overall, treatment and control) with the 90% confidence intervals, but this time under different assumptions (Assumptions 2 and equation 38). The average total effect and its 90% confidence intervals, which are essentially identical to those in Figure 2, are also shown at the bottom. Unlike Assumption 1, the estimated overall ACME is statistically indistinguishable from zero, although it is similar in magnitude (0.313 with the confidence interval of $[-0.021, 0.648]$) and represents about 32.2 percent of the total framing effect. The ACME does not vary as much according to the baseline treatment status, with the estimate being statistically insignificant in either case (0.265, $[-0.077, 0.606]$ for the treatment condition; 0.332, $[-0.092, 0.755]$ for the control). Thus, the results give only weak support to the original conclusion once the assumed causal independence between the mechanisms is relaxed.

This analysis relaxes the independence assumption but still assumes that the interaction between framing effects and the effects of issue importance is homogeneous across participants (equation 38). Because this assumption is rather strong, we examine the robustness of our conclusions to its violation via the proposed sensitivity analysis. The remaining two panels on the left column of Figure 3 show the results. In the first panel, the sharp bounds on the true values of the overall ACME are plotted as functions of the sensitivity parameter σ (thick solid lines), which represents the degree of heterogeneity in the interaction between the issue frames and perceived importance. As discussed in Section 5, observed data imply an upper bound on the possible values of this parameter, which equals the rightmost value of σ in the figure (0.382) and represents the maximal possible violation of equation (38) given the data. The 90% confidence intervals are computed based on the approach of Imbens and Manski (2004) with the bootstrap standard errors and represented by the gray region around the bounds. The point-identified value of ACME under equation (38) is represented by the dashed horizontal line.

The result shows that the lower bound of ACME does not become negative until σ becomes greater than 0.195, or 51% of its largest possible value. This implies that, disregarding the statistical uncertainty in these estimates, the data from Druckman and Nelson's experiment provide some support for the importance mechanism even if we allow for a certain degree of violation of the no interaction assumption. When we completely relax the no interaction assumption and only make the sequential ignorability assumption (Assumption 2), however, the bounds become wide ($[-0.303, 0.922]$) with the 90% confidence interval of $[-0.618, 1.159]$.

The bottom figure presents the result of the same sensitivity analysis using an alternative parameterization. As discussed in Section 5, the $\tilde{R}^2$ parameter here represents the proportion of the total variance of the outcome variable that would be explained if we could take into account the heterogeneity

of the treatment-mediator interaction in the regression model. The upper bound of $\tilde{R}^2$, 0.610, corresponds to the scenario where the residual variation in the outcome variable is completely attributed to the interaction heterogeneity and thus represents the severest possible violation of the no interaction assumption given the observed data. In this alternative formulation, the bounds on the overall ACME include zero only when $\tilde{R}^2$ is greater than 0.159. This implies that we can maintain the conclusion of the original analysis unless the interaction heterogeneity explains more than 15.9% of the total variance of the participants' opinion. In summary, if one ignores estimation uncertainty, the conclusions of Druckman and Nelson (2003) are reasonably robust to the violations of the two assumptions made in their original analysis: the causal independence between alternative mechanisms and the no treatment-mediator interaction effect. However, the generally wide confidence intervals suggest that drawing definite conclusions would require a study with a larger sample size.

6.2 Slothuus (2008)

The framing experiment of Slothuus is also subject to the possibility that the participants' perceived importance of work incentives may be influenced by the content of their considerations about the welfare reform. If this were the case, the importance mechanism is causally dependent on the belief content mechanism. We therefore apply our proposed method to re-estimate the ACME for the importance mechanism relaxing the independence assumption. Again, we use the same set of pretreatment covariates as in Section 4.2. The results are presented in the middle column of Figure 3.

The top panel shows that under the assumptions of sequential ignorability (Assumption 2) and homogeneous interaction (equation 38), the estimated overall ACME for the importance mechanism is about 0.159, representing 14.1 percent of the total effect (1.128). The 90% confidence interval for this estimate does not contain zero ($[0.004, 0.315]$). The estimates for the treatment and control conditions are slightly different (0.102 with $[-0.054, 0.257]$ and 0.208 with $[0.004, 0.411]$, respectively). These estimates are both smaller in magnitude and weaker in statistical significance than those assuming the causal independence between the importance and content mechanisms (see Section 4.2).

The other two panels show the results of the sensitivity analysis. As we gradually relax the homogeneous interaction assumption, the 90% confidence interval for the overall ACME covers zero almost immediately. The lower bound also becomes less than zero when σ is greater than 0.607, or about 24.5 percent of its largest possible value implied by the data (0.987). This indicates that the result of Slothuus's experiment is more sensitive to the heterogeneity in the treatment-mediator interaction than Druckman and Nelson's result, despite the statistical significance of the original point estimate. The result translates to the value of 0.029 in terms of the alternative $\tilde{R}^2$ parameter, suggesting that

the interaction heterogeneity must only explain 2.9 percent of the total variation in the participants' opinion (out of the maximum value of 48.9 percent), so that the bounds on the overall ACME for the importance mechanism contain zero. In summary, once we allow for the dependence among mechanisms, the original conclusions from Slothuus's framing experiment are quite sensitive to the violation of no interaction even under the exogeneity assumptions.

6.3 Brader, Valentino and Suhay (2008)

Finally, we apply our proposed framework to the experiment of Brader et al.. Like the above two experiments, the assumption of independence between the anxiety and perceived harm mechanisms (Assumption 1) is problematic in Brader et al.'s study both from theoretical and empirical perspectives. The right column of Figure 3 shows the results of our analysis using the same pretreatment covariates as in the original study but accommodating the dependence of anxiety on perceived harm.

Unlike the other two examples, the overall ACME for the anxiety mechanism remains statistically significant at the .1 level under Assumption 2 and equation (38) ($[0.030, 0.182]$). However, compared to the estimate under the independence assumption, the estimated overall ACME is substantially smaller (0.106) and represents only 25.3 percent of the total framing effect (0.420). Moreover, the estimated ACME for the treatment condition is even smaller (0.057) and statistically insignificant ($[-0.073, 0.187]$), while the estimate for the control condition is significant (0.123, $[0.042, 0.204]$). The evidence for the anxiety mechanism thus appears much weaker when we allow the participants' perceived harm to influence their anxiety levels.

This conclusion is reinforced when we examine the sensitivity of these estimates to the violation of the homogeneous interaction effect assumption (equation 38). The middle panel shows that the lower bound of the ACME becomes less than zero when the standard deviation of the coefficient of the interaction term is greater than 0.042, or about 24.5 percent of its largest possible value given the data (0.173). This is roughly the same degree of sensitivity as the Slothuus study. The 90% confidence interval overlaps with zero when σ is 0.013 or greater. The bottom panel indicates that these values of σ translate to the $\tilde{R}^2$ values of 0.036 and 0.006, respectively, implying that the heterogeneity in the interaction between the frame and anxiety must only explain 3.6 percent of the total variance in the outcome variable out of its 59.3 percent residual variance. Thus, the evidence for the anxiety mechanism in Brader et al.'s study becomes quite fragile when we allow for its dependence on the participants' perceived harm of immigration.

7 Extension to the New Experimental Designs

So far we have analyzed standard randomized experiments where the treatment is randomized but the mediator is not. In this setup, the exogeneity of the mediator must be assumed instead of guaranteed by the experimental design. This remains an important limitation even though our framework described in Section 5 allows for the existence of post-treatment observed confounders. For example, in Druckman and Nelson's framing experiment, we cannot rule out the possibility that there exists an unobserved post-treatment confounder affecting both the perceived issue importance and opinions about the campaign finance reform, even after taking into account the belief content mechanism. In this section, we extend our method to the two new experimental designs recently proposed by Imai et al. (2012) which avoid the assumption of exogenous mediators. In particular, we first consider the parallel design where two randomized experiments are run in parallel; one experiment employs the standard design where only the treatment is randomized whereas the other experiment randomizes both the treatment and the mediator of interest. We also consider the parallel encouragement design where the manipulation of the mediator is imperfect in the second experiment.

7.1 The Parallel Design

Under the parallel design, we randomly split the sample into two groups and conduct a separate randomized experiment for each group in parallel. One experiment is conducted under the standard design where, after the treatment is randomized and administered, the values of the mediator and the outcome are measured. In the other experiment, both the treatment and mediating variables are randomized and subsequently the values of the outcome variable are recorded. We assume that the values of the potential outcomes remain identical regardless of the experiment to which each unit is assigned. In other words, the potential outcomes are assumed to depend on the values of the treatment and mediator but not how they are realized. Imai et al. (2012) emphasize the importance of this assumption and suggest that the manipulation of the mediator needs to be subtle in order to make sure that it affects the outcome only through the mediator value and has no direct effect.

Formally, let D_i be an indicator variable representing whether unit i is randomly assigned to the first ($D_i = 0$) or second ($D_i = 1$) experiment. The study design implies the following conditional independence,

$$\{Y_i(t, m), M_i(t')\} \quad \perp\!\!\!\perp \quad T_i \mid D_i = 0, \tag{43}$$

$$\{Y_i(t, m)\} \quad \perp\!\!\!\perp \quad \{T_i, M_i\} \mid D_i = 1, \tag{44}$$

for any t, t', m. Note that we do not assume independence between the mediator and potential outcomes in the first experiment. This means that while the manipulated values of the mediator must be random in the second experiment, the parallel design makes no exogeneity assumption on the mediator values that would naturally realize when only the treatment is manipulated as in the first experiment. This contrasts with the scenarios depicted in Figure 1 where observed mediators are assumed to be conditionally exogenous. The parallel design also differs from Figure 1 in that there is no need to determine whether there exist alternative mechanisms that are causally related to the mechanism of interest.

Under the parallel design, the first experiment identifies the average effects of the treatment on the mediator and outcome. That is, both $\mathbb{E}\{M_i(t)\}$ and $\mathbb{E}\{Y_i(t, M_i(t))\}$ are identifiable for $t = 0, 1$. The second experiment, on the other hand, identifies $\mathbb{E}\{Y_i(t, m)\}$ for any t and m. Unfortunately, this is not sufficient to identify the ACME. Following Robins (2003), Imai et al. (2012) show that under the parallel design, the ACME is nonparametrically identified if the assumption of no interaction effect holds. This assumption, which is essentially the same as Assumption 3, can be formally written as,

$$Y_i(1, m) - Y_i(1, m') \quad = \quad Y_i(0, m) - Y_i(0, m'), \tag{45}$$

for any $m \neq m'$. As is the case for Assumption 3, the equality must hold at the unit level rather than in expectation, which makes this assumption unverifiable and unrealistic in most cases. However, without this assumption, the identification power of the parallel design is quite limited. Imai et al. (2012) show that if the outcome variable is unbounded, the ACME cannot be bounded. They also show that even when the outcome and mediator are binary, the sharp bounds on the ACME may often contain zero, failing to identify its sign (see also Sjölander, 2009; Kaufman et al., 2009). Thus, below, we develop a sensitivity analysis for the no interaction effect assumption to assess the consequence of the violation of this key identification assumption.

7.2 The Proposed Methodology for the Parallel Design

Given the setup described above, we develop a sensitivity analysis for inference based upon the following system of linear equations with varying coefficients,

$$M_i(t) \quad = \quad \alpha_2 + \beta_{2i} t + \varepsilon_{2i} \tag{46}$$

$$Y_i(t, m) \quad = \quad \alpha_3 + \beta_{3i} t + \gamma_i m + \kappa_i t m + \varepsilon_{3i}, \tag{47}$$

for any t and m where $\mathbb{E}(\varepsilon_{2i}) = \mathbb{E}(\varepsilon_{3i}) = 0$ without loss of generality. Like the one considered in Section 5, this model allows for arbitrary degrees of heterogeneity across units and does not make any distri-

butional assumption about these effects. Thus, despite the linearity assumption, equations (46) and (47) represent a general class of models that are sufficiently flexible to be useful in a variety of applied research settings.

Under the parallel design, equations (46) and (47) can be fitted separately to the data from the first and second experiments, respectively. The randomization of the treatment in the first experiment ($D_i = 0$) and that of the treatment and mediator in the second experiment ($D_i = 1$) guarantees that the following exogeneity assumptions are satisfied,

$$\mathbb{E}(\varepsilon_{2i} \mid T_i, D_i = 0) \;=\; \mathbb{E}(\varepsilon_{3i} \mid T_i, M_i, D_i = 1) \;=\; 0. \tag{48}$$

In addition, due to the linearity and the binary nature of the treatment, the model implies the following linear relationship between the treatment and the outcome,

$$Y_i(t, M_i(t)) \;=\; \alpha_1 + \beta_{1i} t + \varepsilon_{1i} \tag{49}$$

where $\alpha_1 + \varepsilon_{1i} = \alpha_3 + (\alpha_2 + \varepsilon_{2i})\gamma_i + \varepsilon_{3i}$, $\beta_{1i} = \beta_{3i} + \beta_{2i}\gamma_i + (\alpha_2 + \beta_{2i} + \varepsilon_{2i})\kappa_{2i}$ and $\mathbb{E}(\varepsilon_{1i}) = 0$.

Now, we can rewrite the model given in equations (46) and (47) as,

$$M_i(t) \;=\; \alpha_2 + \beta_2 t + \eta_{2i}(t) \tag{50}$$

$$Y_i(t, m) \;=\; \alpha_3 + \beta_3 t + \gamma m + \kappa t m + \eta_{3i}(t, m) \tag{51}$$

where $\beta_2 = \mathbb{E}(\beta_{2i})$, $\beta_3 = \mathbb{E}(\beta_{3i})$, $\gamma = \mathbb{E}(\gamma_i)$, $\kappa = \mathbb{E}(\kappa_i)$, and $\eta_{2i}(t) = \tilde{\beta}_{2i} t + \varepsilon_{2i}$ and $\eta_{3i}(t, m) = \tilde{\beta}_{3i} t + \tilde{\gamma}_i m + \tilde{\kappa}_i t m + \varepsilon_{3i}$ with $\mathbb{E}(\tilde{\beta}_{2i}) = \mathbb{E}(\tilde{\beta}_{3i}) = \mathbb{E}(\tilde{\gamma}_i) = \mathbb{E}(\tilde{\kappa}_i) = 0$. Thus, under the parallel design, the exogeneity assumption given in equation (48) implies that among the parameters of equations (50) and (51), $(\alpha_2, \alpha_3, \beta_2, \beta_3, \gamma, \kappa)$ are identified.

To develop a sensitivity analysis, we follow the analytical strategy employed in Section 5 and write the average direct effect using the model parameters as follows,

$$\bar{\delta}(t) \;=\; \beta_1 - \bar{\zeta}(1 - t) \tag{52}$$

$$\bar{\zeta}(t) \;=\; \beta_3 + (\alpha_2 + \beta_2 t)\kappa + \rho_t \sigma \sqrt{\mathbb{V}(M_i \mid T_i = t, D_i = 0)} \tag{53}$$

for $t = 0, 1$ where the two sensitivity parameters are $\rho_t = \mathrm{Corr}(M_i(t), \kappa_i)$ and $\sigma = \sqrt{\mathbb{V}(\kappa_i)}$ and other parameters can be consistently estimated from the observed data because other parameters are identifiable under this design. These sensitivity parameters, ρ_t and σ, represent the degree to which

the individual-level treatment-mediator interaction effect is correlated with the mediator of interest and the amount of heterogeneous interaction effect, respectively. Researchers can vary these two sensitivity parameters within their plausible range to assess the sensitivity of their empirical results to the violation of the no-interaction effect assumption under the parallel design.

This sensitivity analysis under the parallel design is essentially identical to that of Section 5 except that there is no need to consider alternative mediators W since M is randomized. Thus, most of the remarks made in Section 5 also apply here. First, the no-interaction effect assumption given in equation (45) implies $\kappa_i = 0$ for all i and thus $\kappa = \rho_t = \sigma = 0$. In this case, both the ACME and the average direct effect are identified as $\bar{\delta}(t) = \beta_1 - \beta_3 = \beta_2\gamma$ and $\bar{\zeta}(t) = \beta_3$, respectively, which equal the usual mediation procedure under sequential ignorability without the treatment-mediator interaction.

Second, one can consider the following homogeneous interaction effect assumption, $Y_i(1, m) - Y_i(0, m) = B_i + Cm$, where C does not vary across units. Under the linear models considered here this assumption implies $\tilde{\kappa}_i = 0$ for all i and thus $\sigma = 0$. Therefore, both the ACME and the average direct effect are identified as $\bar{\delta}(t) = \beta_1 - \beta_3 - (\alpha_2 + \beta_2(1-t))\kappa = \beta_2(\gamma + t\kappa)$ and $\bar{\zeta}(t) = \beta_3 + (\alpha_2 + \beta_2 t)\kappa$, respectively. These formulae agree with the standard mediation procedure under the linear structural modeling with the treatment-mediator interaction term (e.g., Kraemer et al., 2008). Our analysis therefore highlights the implicit assumption made when researchers apply the standard mediation analysis procedure.

Third, if neither of these assumptions holds, then the standard estimates of the ACME and the average direct effect will be biased even under the parallel design where assumptions (43) and (44) are both satisfied. For example, the bias for the average direct effect $\bar{\zeta}(t)$ is equal to $\rho_t\sigma\sqrt{\mathbb{V}(M_i \mid T_i = t, D_i = 0)}$. This implies that $\bar{\zeta}(t)$ will be overestimated if the mediator positively interacts with the treatment for those units who tend to have high mediator values when the treatment status is t (i.e. $\rho_t > 0$). The magnitude of such bias will be large when the degree of heterogeneity for the treatment-mediator interaction effect is high (i.e. σ is large).

Finally, as in Section 5, this sensitivity analysis can be conducted with respect to an alternative sensitivity parameter instead of σ for easier interpretation. Specifically, we can use the proportion of the unexplained or original variance of the outcome that is additionally explained by including the heterogeneity in the treatment-mediator interaction. This quantity is represented by the following coefficients of determination, $R^{2*} = \mathbb{V}(\kappa_i T_i M_i \mid D_i = 1)/\mathbb{V}(\eta_{3i}(T_i, M_i) \mid D_i = 1)$ for the unexplained variance and $\widetilde{R}^2 = \mathbb{V}(\kappa_i T_i M_i \mid D_i = 1)/\mathbb{V}(Y_i \mid D_i = 1)$ for the original variance. For example, R^{2*} represents how much of the observed variance in Y_i can be explained by the inclusion of the term $\kappa_i T_i M_i$ in the regression model. Then, it can be shown that σ can be alternatively expressed as a function of

each of these coefficients of determination,

$$\sigma \;=\; \sqrt{\frac{\mathbb{V}(\eta_{3i}(T_i, M_i) \mid D_i = 1)R^{2^*}}{\mathbb{E}(T_i M_i^2 \mid D_i = 1)}} \;=\; \sqrt{\frac{\mathbb{V}(Y_i \mid D_i = 1)\widetilde{R}^2}{\mathbb{E}(T_i M_i^2 \mid D_i = 1)}}. \tag{54}$$

Thus, we can conduct the sensitivity analysis with the sensitivity parameters R^{2^*} and $\widetilde{R}^2$. From this expression, it is immediate that the upper bound of σ results when the heterogeneity in the treatment-mediator interaction explains all the unexplained variance, i.e., $\sigma \leq \sqrt{\mathbb{V}(\eta_{3i}(T_i, M_i) \mid D_i = 1)/\mathbb{E}(T_i M_i^2 \mid D_i = 1)}$.

7.3 The Parallel Encouragement Design

The sensitivity analysis developed above can be extended to the "parallel encouragement design" of Imai et al. (2012) where the manipulation of the mediator in the second experiment of the parallel design is imperfect. This situation is more realistic for psychological studies such as framing experiments, because manipulating psychological mechanisms is likely to be imperfect at best even with clever use of intervention techniques.

Under this design, it is assumed that the randomized manipulation of the mediator monotonically affects the mediator and this manipulation affects the outcome only through the realized value of the mediator. Then, the manipulation Z_i can be used as an instrumental variable and we can fit the two-stage least squares regression model. That is, equation (46) can be replaced with the following equation,

$$M_i(t, z) \;=\; \alpha_2 + \beta_{2i} t + \lambda_i z + \theta_i t z + \varepsilon_{2i} \tag{55}$$

whereas equation (47) remains identical. Under this two-stage least squares model, it is straightforward to show that the average direct effect and ACME, $\bar{\zeta}(t, z) = \mathbb{E}\{Y_i(1, M_i(t, z)) - Y_i(0, M_i(t, z))\}$ and $\bar{\delta}(t, z) = \mathbb{E}\{Y_i(t, M_i(1, z)) - Y_i(t, M_i(0, z))\}$, equal the following expressions,

$$\bar{\zeta}(t, z) \;=\; \beta_3 + (\alpha_2 + \beta_2 t + \lambda z + \theta t z)\kappa + \rho_{t,z}\sigma\sqrt{\mathbb{V}(M_i \mid T_i = t, Z_i = z)}, \tag{56}$$

$$\bar{\delta}(t, z) \;=\; \bar{\tau} - \beta_3 - (\alpha_2 + \beta_2(1 - t) + \lambda z + \theta t z)\kappa - \rho_{1-t,z}\sigma\sqrt{\mathbb{V}(M_i \mid T_i = 1 - t, Z_i = z)}, \tag{57}$$

where $\lambda = \mathbb{E}(\lambda_i)$, $\theta = \mathbb{E}(\theta_i)$ (which are both identified due to the exogeneity of T_i and Z_i) and $\rho_{t,z} = \mathrm{Corr}(M_i(t, z), \kappa_i)$ for $t \in \{0, 1\}$ and $z \in \mathcal{Z}$ (the support of Z_i). The average total effect, $\bar{\tau}$ in equation (57), can be consistently estimated by regressing Y_i on T_i for the subsample who received the manipulation of the treatment alone, i.e., $Z_i = 0$. Then, the sensitivity and bounds analyses can be conducted by estimating the identifiable parameters via the two-stage least squares method and varying

$\rho_{t,z}$ and σ within the ranges of plausible values.

8 Concluding Remarks and Suggestions for Applied Researchers

In this paper, we have shown how to conduct causal mediation analysis in the presence of multiple mediators. The proposed methodology can be applied even when alternative mechanisms are causally related to each other so long as both the treatment and mediating variables can be assumed to be exogenous conditional on a set of observed confounders. We also show that our methodology can be applied to new experimental designs where these exogeneity conditions can be ensured by experimenters. Finally, we conclude this paper by offering a list of practical suggestions for applied researchers who wish to study multiple causal mechanisms.

- Identify a list of alternative causal mechanisms that are causally prior to the mechanism of your interest and measure mediators that represent them.
- Consider theoretically whether the identified alternative mechanisms are causally related to the causal mechanisms of interest and test empirically whether they are statistically dependent on each other even after adjusting for the treatment and pre-treatment covariates.
- If alternative mechanisms are causally independent of one another, apply the standard causal mediation analysis and conduct sensitivity analysis for the possible existence of unobserved pre-treatment confounders.
- If alternative mechanisms are causally related to each other, apply the mediation analysis that directly accounts for multiple mediators and conduct sensitivity analysis with respect to the existence of treatment-mediator interaction effects.
- Whenever possible, utilize research designs, experimental or observational, where the exogeneity of mediator is credible and apply sensitivity analysis for the no treatment-mediator interaction assumption.

A Mathematical Appendix

A.1 Proof of Nonparametric Identification of Average Causal Mediation Effects under Assumption 1

The proof is a straightforward extension of that of Theorem 1 in Imai et al. (2010c). We first consider the identification of $\bar{\delta}_i^M(t)$, the average causal mediation effect with respect to M_i. Note that equation (18) implies the following conditional independence:

$$Y_i(t, m, W_i(t)) \quad \perp\!\!\!\perp \quad T_i \;\mid\; M_i(t') = m', X_i = x. \tag{58}$$

for all $t, t' = 0, 1$, m, m', and x. Now, for any t, t', we have,

$$\begin{aligned}
& \mathbb{E}(Y_i(t, M_i(t'), W_i(t)) \mid X_i = x) \\
= & \int \mathbb{E}(Y_i(t, m, W_i(t)) \mid M_i(t') = m, X_i = x)\, dF_{M_i(t')|X_i=x}(m) && (59) \\
= & \int \mathbb{E}(Y_i(t, m, W_i(t)) \mid M_i(t') = m, T_i = t', X_i = x)\, dF_{M_i(t')|X_i=x}(m) && (60) \\
= & \int \mathbb{E}(Y_i(t, m, W_i(t)) \mid T_i = t', X_i = x)\, dF_{M_i(t')|X_i=x}(m) && (61) \\
= & \int \mathbb{E}(Y_i(t, m, W_i(t)) \mid T_i = t, X_i = x)\, dF_{M_i(t')|T_i=t',X_i=x}(m) && (62) \\
= & \int \mathbb{E}(Y_i(t, m, W_i(t)) \mid M_i(t) = m, T_i = t, X_i = x)\, dF_{M_i(t')|T_i=t',X_i=x}(m) && (63) \\
= & \int \mathbb{E}(Y_i \mid M_i = m, T_i = t, X_i = x)\, dF_{M_i(t')|T_i=t',X_i=x}(m) && (64) \\
= & \int \mathbb{E}(Y_i \mid M_i = m, T_i = t, X_i = x)\, dF_{M_i|T_i=t',X_i=x}(m), && (65)
\end{aligned}$$

where the second equality follows from equation (58), equation (18) is used to establish the third and fifth equalities, equation (17) is used to establish the fourth and last equalities, and the sixth equality follows from the fact that $M_i = M_i(T_i)$ and $Y_i = Y_i(T_i, M_i(T_i), W_i(T_i))$. This implies that $\bar{\delta}^M(t)$ can be identified by the identical expression as Theorem 1 in Imai et al. (2010c). Furthermore, the same proof applies to $\bar{\delta}^W(t)$ by considering the identification of $\mathbb{E}(Y_i(t, M_i(t), W_i(t')) \mid X_i = x)$ using equation (19) instead of (18). Finally, $\bar{\zeta}(t, t')$ is also identified because $\bar{\zeta}(t, t') = \bar{\tau} - \bar{\delta}^M(t') - \bar{\delta}^W(t)$ and $\bar{\tau}$ is identified under equation (17). □

A.2 Proof of the Parametric Identification of Average Causal Mediation Effects with the Path Analysis under Assumption 1

The proof follows the similar argument as Theorem 2 of Imai et al. (2010c). First, note that the coefficients in equations (20), (21) and (22) are all identified under Assumption 1. Next, note that given this linear structural equations model, equation (65) can be written as,

$$
\begin{aligned}
& \mathbb{E}(Y_i(t, M_i(t'), W_i(t)) \mid X_i = x) \\
= & \int \mathbb{E}(Y_i \mid M_i = m, T_i = t, X_i = x)\, dF_{M_i|T_i=t',X_i=x}(m) && (66) \\
= & \int\int \left(\alpha_3 + \beta_3 t + \gamma m + \theta^\top w + \xi_3^\top x\right)\, dF_{W_i|M_i=m,T_i=t,X_i=x}(w)\, dF_{M_i|T_i=t',X_i=x}(m) && (67) \\
= & \int\int \left(\alpha_3 + \beta_3 t + \gamma m + \theta^\top w + \xi_3^\top x\right)\, dF_{W_i|T_i=t,X_i=x}(w)\, dF_{M_i|T_i=t',X_i=x}(m) && (68) \\
= & \alpha_3 + \beta_3 t + \gamma(\alpha_M + \beta_M t' + \xi_M^\top x) + \theta^\top(\alpha_W + \beta_W t + \xi_W^\top x) + \xi_3^\top x && (69)
\end{aligned}
$$

where the third equality holds because equation (18) implies $M_i \perp\!\!\!\perp W_i \mid T_i = t, X_i = x$ under the linear structural equations model in equations (20), (21) and (22). Finally, we also have

$$
\mathbb{E}(Y_i(t, M_i(t), W_i(t)) \mid X_i = x) \;=\; \alpha_3 + \beta_3 t + \gamma(\alpha_M + \beta_M t + \xi_M^\top x) + \theta^\top(\alpha_W + \beta_W t + \xi_W^\top x) + \xi_3^\top x. \quad (70)
$$

Therefore, $\bar{\delta}^M(t) = \beta_M \gamma$. The same argument applies to $\bar{\delta}^W(t)$ and it is identified as $\beta_W \theta$. □

A.3 Proof of the Identification of the Average Direct Effect Given ρ_t and σ

We begin by expressing the average direct effect defined in equation (24) using the parameters of the model given in equations (30) and (31),

$$
\begin{aligned}
\bar{\zeta}(t) &= \mathbb{E}\{\beta_{3i} + \kappa_i(\alpha_2 + \beta_{2i} t + \xi_{2i}^\top W_i(t) + \mu_{2i}^\top t W_i(t) + \lambda_{2i}^\top x + \varepsilon_{2i}) + \xi_{3i}^\top(W_i(1) - W_i(0)) + \mu_{3i}^\top W_i(1)\} \\
&= \beta_3 + \kappa \mathbb{E}(M_i \mid T_i = t) + \rho_t \sigma \sqrt{\mathbb{V}(M_i \mid T_i = t)} + \mathbb{E}\{(\xi_{3i} + \mu_{3i})^\top W_i(1) - \xi_{3i}^\top W_i(0)\}, \qquad (71)
\end{aligned}
$$

where the two conditional moments of M_i, $\mathbb{E}(M_i \mid T_i = t)$ and $\mathbb{V}(M_i \mid T_i = t)$, can be consistently estimated using their sample counterparts. Now, note that the last term of equation (71) can be written as,

$$
\begin{aligned}
\mathbb{E}\{(\xi_{3i} + \mu_{3i})^\top W_i(1) - \xi_{3i}^\top W_i(0)\} &= \mathbb{E}\{(\xi_{3i} + \mu_{3i})^\top W_i(1) \mid T_i = 1\} - \mathbb{E}(\xi_{3i}^\top W_i(0)) && (72) \\
&= \mathbb{E}\{(\xi_{3i} + \mu_{3i})^\top \mid T_i = 1\}\mathbb{E}(W_i(1) \mid T_i = 1) - \mathbb{E}(\xi_{3i}^\top W_i(0)) && (73) \\
&= (\xi_3 + \mu_3)^\top \mathbb{E}(W_i \mid T_i = 1) - \mathbb{E}(\xi_{3i}^\top W_i(0)), && (74)
\end{aligned}
$$

where the equalities follow from equation (26) and the fact that each element of the coefficient vector $\xi_{3i} + \mu_{3i}$ is conditionally independent of $W_i(1)$ given $T_i = 1$ under Assumption 2. The latter holds because for any j and m, $\xi_{3ij} + \mu_{3ij} = Y_i(1, m, w) - Y_i(1, m, w')$, where ξ_{3ij} and μ_{3ij} denote the jth elements of ξ_{3i} and μ_{3i}, respectively, $w = (w_1, ..., w_j, ..., w_J)^\top$, and $w' = (w_1, ..., w_j - 1, ..., w_J)^\top$. Likewise, we have,

$$\mathbb{E}\{(\xi_{3i} + \mu_{3i})^\top W_i(1) - \xi_{3i}^\top W_i(0)\} \quad = \quad \mathbb{E}((\xi_{3i} + \mu_{3i})^\top W_i(1)) - \xi_3^\top \mathbb{E}(W_i \mid T_i = 0), \tag{75}$$

since $\xi_{3ij} = Y_i(0, m, w) - Y_i(0, m, w')$ for any $m \in \mathcal{M}$ and $j \in \{1, ..., J\}$. Together with equation (74), we obtain,

$$\mathbb{E}\{(\xi_{3i} + \mu_{3i})^\top W_i(1) - \xi_{3i}^\top W_i(0)\} \quad = \quad (\xi_3 + \mu_3)^\top \mathbb{E}(W_i \mid T_i = 1) - \xi_3^\top \mathbb{E}(W_i \mid T_i = 0), \tag{76}$$

which implies that the final term of equation (71), $\mathbb{E}\{(\xi_{3i} + \mu_{3i})^\top W_i(1) - \xi_{3i} W_i(0)\}$, is identified. Therefore, the average direct effect is identified given the two sensitivity parameters, ρ_t and σ, and so is the average causal mediation effect, which can be obtained by subtracting the average direct effect from the average total effect. □

A.4 Proof of the Identification of the Causal Mediation Effect Specific to the Path $T \to W \to Y$

First, note that the population average of the path-specific effect in equation (41) can be written using the model parameters given in equations (30) and (31) as,

$$\bar{\chi}(t) \quad = \quad \mathbb{E}\{(\xi_{3i} + t\mu_{3i})^\top (W_i(1) - W_i(0))\} \tag{77}$$

Now, following the same argument as in Appendix A.3, we can show that under Assumption 2 this expectation can be written as a function of identified model parameters,

$$\mathbb{E}\{(\xi_{3i} + t\mu_{3i})^\top (W_i(1) - W_i(0))\} \quad = \quad (\xi_3 + t\mu_3) \left\{\mathbb{E}(W_i \mid T_i = 1) - \mathbb{E}(W_i \mid T_i = 0)\right\}. \tag{78}$$

□

References

Albert, J. M. and Nelson, S. (2011). Generalized causal mediation analysis. Biometrics **67**, 3, 1028–1038.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, 357–363, Edinburgh, Scotland. Morgan Kaufmann.

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology **51**, 6, 1173–1182.

Brader, T., Valentino, N., and Suhay, E. (2008). What triggers public opposition to immigration? anxiety, group cues, and immigration threat. American Journal of Political Science **52**, 4, 959–978.

Bullock, J., Green, D., and Ha, S. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). Journal of Personality and Social Psychology **98**, 4, 550–558.

Callaghan, K. and Schnell, F., eds. (2005). Framing American Politics. University of Pittsburgh Press, Pittsburgh, OR.

Chong, D. and Druckman, J. N. (2007). A theory of framing and opinion formation in competitive elite environments. Journal of Communication **57**, 99–118.

Druckman, J. N. and Nelson, K. R. (2003). Framing and deliberation: How citizens' conversations limit elite influence. American Journal of Political Science **47**, 4, 729–745.

Glynn, A. N. (2012). The product and difference fallacies for indirect effects. American Journal of Political Science Forthcoming.

Hafeman, D. (2008). Opening the Black Box: A Reassessment of Mediation from a Counterfactual Perspective. Ph.D. thesis, Columnbia University, New York.

Holland, P. W. (1986). Statistics and causal inference (with discussion). Journal of the American Statistical Association **81**, 945–960.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. Psychological Methods **15**, 4, 309–334.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2010b). Advances in Social Science Research Using R (ed. H. D. Vinod), chap. Causal Mediation Analysis Using R, 129–154. Lecture Notes in Statistics. Springer, New York.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (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, K., Keele, L., and Yamamoto, T. (2010c). Identification, inference, and sensitivity analysis for causal mediation effects. Statistical Science **25**, 1, 51–71.

Imai, K., Tingley, D., and Yamamoto, T. (2012). Experimental designs for identifying causal mechanisms (with discussions). Journal of the Royal Statistical Society, Series A (Statistics in Society) Forthcoming.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. Econometrica **72**, 6, 1845–1857.

Isbell, L. and Ottati, V. (2002). The emotional voter: Effects of episodic affective reactions on candidate evaluation. In V. Ottati, R. Tindale, J. Edwards, F. Bryant, L. Heath, D. O'Connell, Y. Suarez-Balcazar, and E. Posavac, eds., The Social Psychology of Politics. Social Psychological Application to Social Issues, Vol. 5, 55–74. Kluwer, New York.

Iyengar, S. (1991). Is Anyone Responsible? The University of Chicago Press, Chicago.

Kaufman, S., Kaufman, J. S., and MacLehose, R. F. (2009). Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. Journal of Statistical Planning and Inference **139**, 3473–3487.

Kinder, D. R. and Sanders, L. M. (1990). Mimicking political debate with survey questions: The case of white opinion on affirmative action for blacks. Social Cognition **8**, 1, 73–103.

Kraemer, H. C., Kiernan, M., Essex, M., and Kupfer, D. J. (2008). How and why criteria definig moderators and mediators differ between the Baron & Kenny and MacArthur approaches. Health Psychology **27**, 2, S101–S108.

Miller, J. M. (2007). Examining the mediators of agenda setting: A new experimental paradigm reveals the role of emotions. Political Psychology **28**, 6, 689–717.

Nelson, T. E., Clawson, R. A., and Oxley, Z. M. (1997). Media framing of a civil liberties conflict and its effect on tolerance. American Political Science Review **91**, 3, 567–583.

Nelson, T. E. and Kinder, D. R. (1996). Issue framing and goup-centrism in american public opinion. Journal of Politics **58**, 1055–1078.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). Statistical Science **5**, 465–480.

Pearl, J. (2001). Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, 411–420, San Francisco, CA. Morgan Kaufmann.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of direct causal effects. Epidemiology **17**, 3, 276–284.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods: Application to control of the healthy worker survivor effect. Mathematical Modeling **7**, 1393–1512.

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

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology **3**, 2, 143–155.

Robins, J. M. and Richardson, T. (2010). Alternative graphical causal models and the identification of direct effects. In P. Shrout, K. Keyes, and K. Omstein, eds., Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures. Oxford University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology **66**, 688–701.

Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. Statistics in Medicine **28**, 4, 558–571.

Slothuus, R. (2008). More than weighting cognitive importance: A dual-process model of issue framing effects. Political Psychology **29**, 1, 1–28.

Taylor, A. B., MacKinnon, D. P., and Tein, J.-Y. (2008). Tests of the three-path mediated effect. Organizational Research Methods **11**, 2, 241–269.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2011). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. Tech. rep., Harvard University School of Public Health, Cambridge, MA.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science **211**, 453–458.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. Epidemiology **20**, 1, 18–26.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology **21**, 4, 540–551.

Zaller, J. (1992). The Nature and Origins of Mass Opinion. Cambridge University Press, New York.