Comparative Causal Mediation and Relaxing the Assumption of No Mediator-Outcome Confounding: An Application to International Law and Audience Costs

Kirk Bansak*

April 2018

For comments and questions, please e-mail: kbansak@stanford.edu

Abstract

Experiments often include multiple treatments, with the primary goal to compare the causal effects of multiple treatments. While comparing the magnitudes of the average treatment effects (ATEs) is straightforward, there exist few methods to systematically compare the causal anatomies of each treatment—that is, the collection of causal mechanisms underlying each treatment’s total effect—in order to understand the sources of their relative magnitudes. This study introduces a framework for comparing the causal anatomies of multiple treatments through the use of causal mediation analysis. The study proposes a set of comparative causal mediation estimands that compare the mediation effects of different treatments via a common mediator. It derives the properties of a set of estimators, which are shown to be consistent (or conservative) under a set of assumptions that do not require the absence of unobserved confounding of the mediator-outcome relationship, which is a strong and nonrefutable assumption that must be made for consistent estimation of individual causal mediation effects. To illustrate the method, this study presents an original application investigating the effect of international legality on the domestic political costs that democratic governments suffer for violating foreign policy commitments.

Keywords: causal mediation, causal inference, experimental design

*Department of Political Science, Stanford University, 616 Serra Street, Stanford, CA 94305-6044.
I thank Avidit Acharya, Justin Grimmer, Jens Hainmueller, Andy Hall, Kosuke Imai, Hye-Sung Kim, Ken Scheve, Mike Tomz, and Teppei Yamamoto for helpful comments.
I. Introduction

Causal mediation analysis aims to open the “black box of causality,” offering the opportunity to explore how and why certain treatment effects occur in addition to simply detecting the existence of those effects. Estimation of causal mediation effects, which are effects transmitted via intermediary variables called mediators, has traditionally been implemented using the parametric structural equations model (SEM) framework (Baron and Kenny, 1986). More recent years have seen important advances in the formalization, generalization, and estimation of causal mediation effects within the potential outcomes framework (Robins and Greenland, 1992; Albert, 2008; Imai et al., 2010a, 2011a,b) and both parametric and nonparametric SEM frameworks (Pearl, 2001; VanderWeele, 2009). Causal mediation analysis is often used in experimental research. In the most commonly used “single-experiment design,” the treatment variable is randomized and the mediator(s) observed.

Another trend in experimental research is the design of more complex experiments featuring multiple treatment arms. As knowledge and empirical results have accumulated in various academic sub-fields and in specific program evaluation contexts, experimental research questions have evolved in ways that require evaluating multiple related treatments. Instead of simply testing the effects of single treatments, often of primary interest are the empirical and theoretical differences between the effects of multiple treatments. Indeed, scientific and social scientific theories usually entail causal mechanisms, and hypotheses on why one treatment should have a larger effect than another tend to be based on presumed mechanism(s) through which each treatment propagates its effect. In addition, program and policy evaluation often involves comparing the efficacy of multiple possible interventions and determining how to most efficiently target different interventions to different contexts. In both research settings, richer insights can be gained from comparing different treatments’ causal anatomies—that is, the ensemble of causal mechanisms that endow each treatment with its effect.

This study presents a new framework for comparing the causal anatomies of multiple treatments through the use of causal mediation analysis. It proposes a novel set of comparative causal mediation (CCM) estimands that compare the mediation effects of different treatments via a common mediator. In addition, the value of the method is enhanced by the fact that, as this study shows, these CCM estimands can be estimated under fewer threats to internal validity than individual causal mediation effects. Specifically, consistent estimation of individual causal mediation effects requires the strong and nonrefutable assumption of no unobserved confounding of the mediator-outcome relationship. In contrast, this study derives the properties of a set of estimators for the CCM estimands and shows these estimators to be consistent (or conservative) under assumptions that do not require the absence of unobserved confounding of the mediator-outcome relationship. The study also derives the finite-sample properties of these estimators, showing how adjustments can be made to improve their central tendencies in small samples.

The estimators are easy to understand and implement, thereby providing researchers with...
a simple, reliable, and systematic method of comparing, discovering, and testing the causal mechanism differences between multiple treatments. An original application, investigating the effect of international legality on the domestic political costs that democratic governments suffer for violating foreign policy commitments, is presented to illustrate the method. Software is also offered to implement the method.

The remainder of this study is organized as follows. Section II provides motivation and explains the value, in both theoretical and policy contexts, for comparing the causal mediation effects of multiple treatments. Section III formally introduces the new CCM estimands. Section IV then presents an estimation framework, describing the assumptions and methods under which the CCM estimands can be estimated consistently. Section V presents simulations to illustrate these properties of the estimation framework. Section VI then describes how these properties change—namely, how the CCM estimands can be estimated conservatively but no longer consistently—under a relaxation of the assumptions. Section VII presents the original application illustrating the CCM method. Section VIII concludes.

II. Motivation for Comparing Causal Mediation Effects

As research agendas in both academic and policy contexts advance, experimental research often involves evaluating distinct but conceptually or administratively related treatments. Comparing the magnitudes of the average treatment effects (ATEs) of multiple related treatments is a valuable first step, but simple ATE comparisons provide little insight into the specific processes by which one treatment achieves a larger effect over another—and, alternatively, whether equally sized treatments are truly interchangeable or instead work via different channels. In research contexts where multiple related treatments are under study, richer insights can be gained from comparing the related treatments’ causal anatomies. Such investigations can be useful for making theoretical discoveries, as well as program evaluation and policy targeting.

A. Testing Theories

While experimental tests of social scientific theories often begin and end with estimation of ATEs, those theories almost always also involve causal mechanisms or channels that remain untested. And in research contexts involving multiple related treatments, theories on why one treatment should have a larger effect than another are based on the presumed mechanism(s) through which each treatment propagates its effect.

As one example, consider the recent accumulation of experimental evidence in the political science literature on “audience costs” (for a brief review, see [Hyde, 2015]). Audience costs refer to the electoral costs to politicians (i.e. punishment by voters) for breaking policy commitments, and the past ten years have seen a deluge of survey experiments providing evidence that voters do, indeed, tend to punish policymakers for reneging on foreign policy commitments (e.g. [Tomz, 2007], [McGillivray and Smith, 2000], [Chaudoin, 2014], [Chilton, 2015], [Hyde, 2015]). These many studies have differed greatly, however, not only in terms of their foreign policy
contexts (e.g. security scenarios, international economic scenarios, etc.) but also in terms of the specific nature of the foreign policy commitment (e.g. informal, legal, etc.). One may then wonder whether and why the nature of such a commitment might affect the strength of the audience cost effect—for instance, through what mechanism may a legalized foreign policy commitment gain audience cost strength over an informal commitment. This is a question that can be answered by comparing the causal anatomies of different audience cost treatments, and an original application is presented in detail later in section VII to explore this issue.

Another example exists in the well-developed literature on party cues in American politics, which includes a wealth of experimental studies that investigate party cue effects on voter attitudes and behavior. As this literature has highlighted, not all party cues are made equal; instead, there are various types of party cues, with empirical evidence suggesting that some work more forcefully than others. For instance, there is some experimental evidence that out-party cues may, in fact, be more influential than in-party cues (Aaroe 2012; Arceneaux and Kolodny 2009; Slothuus and de Vreese 2010; Goren et al. 2009; Nicholson 2012). Yet while there is evidence on the relative magnitudes of these two types of cues, the reasons or precise mechanisms by which out-party cue effects exceed those of in-party cues have been theorized but not tested in a disciplined manner. Doing so would involve a rigorous method for comparing the mechanisms underlying each set of party cues.

B. **Program Evaluation and Policy Targeting**

Comparing the causal anatomies of related treatments also offers great value in the policy and program evaluation context, where multiple related treatments are often investigated in individual studies. The basic goal in this context is to identify, often through experimental studies, which policy interventions are effective in achieving certain economic, social, political, or other outcomes. Ideally, the effectiveness of any preferred policy intervention should be generalizable across time and different localities. However, because of constraints on resources, as well as logistical and administrative realities, the execution of experimental studies is often restricted to short periods of time and small subsets of locations. This results in empirical findings with great internal validity in their localized context but potentially questionable external validity, limiting the researcher’s confidence on whether the treatment found to be most effective in the study would actually be the most effective one more widely or in a full-scale deployment.

Ultimately, the goal of such studies is to enable policymakers (or other actors) to select treatments that will be effective beyond the initial study locality—that is, to help policymakers efficiently target their policy interventions. This requires evaluating how different treatments are likely to perform in different contexts. One important means of doing so is developing a comprehensive understanding of the mechanisms underlying different treatments,

---

1 Party cues are public signals from political parties that associate a party with particular candidates or policy positions, thereby affecting the attractiveness of those candidates or positions for voters who have partisan orientations.
as a treatment’s performance in a particular context depends upon whether or not its primary mechanism(s) can properly operate in that context.

For instance, consider an experimental study on job training programs, aimed at finding employment for lower-income adults. Imagine the study is implemented in a handful of towns and involves multiple training programs under consideration; for simplicity, assume two training programs (i.e. two treatments and a control condition of no training). A preliminary analysis of the results may reveal that both programs have roughly equal-sized effects on employment. A superficial interpretation of these results would then be that the two programs are interchangeable and hence, upon launch of a follow-up large-scale policy intervention, the choice of which specific training program to roll out could be made arbitrarily or based on other criteria. However, given that the experimental results of the study were based on a single sample in a single snapshot in time, such a decision-making rule could be severely sub-optimal.

Instead, to enable efficient policy targeting, it would be useful to investigate the causal mechanism differences between the two job training programs. While their ATEs may have been roughly equal, it is possible they achieved their positive effects on employment via different channels. One program may have achieved its primary effect by increasing the job search motivation of its participants, while the other may have achieved its primary effect by helping its participants to develop specific skills. If equipped with such knowledge, policymakers would be in a much better position to make optimal decisions on which job training program to introduce in different localities. For instance, certain local economies may have a low supply of low-skill jobs and a relatively higher supply of skilled jobs, and thus the training program working primarily through the skill-development channel would be preferred in this locality. In other words, knowledge of the relative causal anatomies of the two training programs would enable more insightful comparative evaluation and ultimately more optimal policy targeting.

III. Comparative Causal Mediation (CCM) Estimands

As described above, much is to be gained from both a theoretical and practical policy standpoint from evaluating the differences in causal anatomy between treatments. This section introduces a set of comparative causal mediation estimands that allow for such an evaluation.

Let $T$ denote a binary treatment variable, $Y$ an outcome variable, and $M$ an intermediary variable that is affected by $T$ and that affects $Y$. Causal mediation effects refer to the average effect of $T$ on $Y$ transmitted via the mediator $M$. This is often termed the natural indirect effect or, in the potential outcomes approach, the average causal mediation effect (ACME). To provide a more formal definition, let $Y(t,m)$ denote the potential outcome for $Y$ given that the treatment $T$ and the mediator $M$ equal $t$ and $m$ respectively, and let $M(t)$ denote the potential value for $M$ given that $T$ equals $t$. Following the potential outcomes approach to causal mediation analysis presented by Imai et al. (2010b), the ACME for a single binary treatment is defined formally as $\kappa(t) = E[Y(t, M(1)) - Y(t, M(0))]$. Note that the ACME is a function of $t$, though in the case of no interaction between the treatment and mediator, the
value of the ACME is the same for $t = 0, 1$.

This study deals with a context in which there are multiple treatments and the researcher is interested in comparing the extent to which those different treatments transmit their effects via a common mediator. For simplicity and conceptual clarity, consider a three-level experimental design that involves a true control condition and two different mutually exclusive treatments. The two treatments may be qualitatively different or one may be a scaled up version of the other. Furthermore, there is a single mediator of interest. It may be the case that multiple mediators have been measured in the experiment, but the estimands of interest will be applied within the context of a single mediator at a time.

Let $T_1$ and $T_2$ denote two mutually exclusive binary treatments and $M$ denote a common mediator. Now define the potential outcomes $Y(t_1, t_2, m)$ and $M(t_1, t_2)$, which allows for defining a separate $ACME_j$ and $ATE_j$ for each treatment $T_j$ as follows:

\[
ACME_1 = \kappa_1(t_1) = E[Y(t_1, 0, M(1, 0)) - Y(t_1, 0, M(0, 0))] \tag{1}
\]
\[
ATE_1 = \tau_1 = E[Y(1, 0, M(1, 0)) - Y(0, 0, M(0, 0))] \tag{2}
\]
\[
ACME_2 = \kappa_2(t_2) = E[Y(0, t_2, M(0, 1)) - Y(0, t_2, M(0, 0))] \tag{3}
\]
\[
ATE_2 = \tau_2 = E[Y(0, 1, M(0, 1)) - Y(0, 0, M(0, 0))] \tag{4}
\]

Note that all effects ($ACME$s and $ATE$s) are referenced against a control condition. In the control condition $T_1 = T_2 = 0$, in the first treatment condition $T_1 = 1$ and $T_2 = 0$, and in the second treatment condition $T_1 = 0$ and $T_2 = 1$.

As will be shown, in spite of the strong assumptions required for the identification of any single ACME, a weaker set of assumptions—which, notably, does not contain the usual assumption of no unobserved confounding of the mediator-outcome relationship—will allow for consistent or conservative estimation of the following two comparative causal mediation (CCM) estimands of interest.

**Definition 1**: Define the estimands of interest as follows:

\[
\text{Estimand 1 : } \frac{ACME_2}{ACME_1} = \frac{\kappa_2(t_2)}{\kappa_1(t_1)} \quad \text{Estimand 2 : } \left(\frac{ACME_2}{ATE_2}\right) = \left(\frac{\kappa_2(t_2)}{\tau_2}\right) \left(\frac{\kappa_1(t_1)}{\tau_1}\right)
\]

The first estimand measures the extent to which treatment 2 has a greater causal mediation effect than treatment 1 in terms of the mediator of interest. This allows for testing whether

---

2It is worth explicitly noting that the method presented in this study does not apply to comparing the effects of a single treatment transmitted via different mediators. In contrast to the method presented in this study, trying to compare the effects transmitted via multiple mediators would compound the threat to internal validity, as the problem of confounding is likely to affect different mediators to different extents and in ways that cannot be measured or tested. As a separate issue, there is also a possibility of causal connections between the mediators, further threatening clean identification and obscuring what is even being measured or estimated. Guidance on how to handle these issues, which are not covered in this study, can be found in Imai and Yamamoto (2013) and Daniel et al. (2015).
one treatment has a stronger absolute effect transmitted through the mediator of interest, relative to another treatment. For instance, in the job training policy evaluation described earlier, it could measure whether one program had a stronger effect transmitted via job search motivation than the other program. In contrast, the second estimand measures the extent to which treatment 2 has a greater proportion of its total effect transmitted through the mediator of interest, relative to treatment 1. This allows for testing the extent to which the mediator is more important to the overall causal anatomy of one treatment versus another. For instance, this could measure whether effect transmission via job search motivation comprised a larger proportion of the overall effect of one job training program versus the other in the example above. Table 1 summarizes the general research questions related to each CCM estimand.

Table 1: General Research Questions Related to Each CCM Estimand

| Estimand 1 | Does $T_2$ exhibit stronger effect transmission via mediator $M$ than $T_1$ does? Does the second treatment have a larger mediated effect in absolute terms? |
|------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|            | $H_0 : \frac{ACME_2}{ACME_1} = 1$ $H_a : \frac{ACME_2}{ACME_1} > 1$ |

| Estimand 2 | Does effect transmission via mediator $M$ make up a larger proportion of $ATE_2$ relative to $ATE_1$? Is $M$ more important for $ATE_2$ than $ATE_1$? |
|------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|            | $H_0 : \left(\frac{ACME_2}{ATE_2}\right) = 1$ $H_a : \left(\frac{ACME_2}{ATE_2}\right) > 1$ |

Which of the two estimands is of interest will depend upon the empirical and theoretical goals of a particular research project. When the researcher’s main goal is to identify which treatment has the strongest absolute effect transmitted via a specific causal channel, the first estimand is likely to be of primary interest. The case of evaluating different job training programs, as presented earlier, provides an example. From the standpoint of optimal policy implementation, the researcher may choose to focus on one specific causal channel, prioritizing transmission of the causal effect via that channel and discounting transmission via other channels. For instance, if the researcher knows that the training programs under consideration will, in the post-evaluation period, be rolled out in target areas where increasing job-search motivation is unlikely to be an effective method of increasing employment (e.g. in local economies with a low supply of low-skill jobs), then it makes sense for the researcher to prioritize the skill-development causal channel. In other words, the researcher’s goal should be to identify which job training program leads to the largest increase in employment specifically via the skill-development channel, regardless of the magnitude of the effect transmitted via the channel of job-search motivation and perhaps even regardless of the relative magnitudes of pro-
grams’ overall ATEs. In that case, the researcher’s goal would be achieved by investigating the first CCM estimand, which would measure how much larger one treatment’s skill-development causal channel is than that of the alternative treatment(s).

If, instead, the researcher is interested in better understanding multiple treatments’ relative causal anatomies more generally, then both the first and second CCM estimands should be of interest. Considering both estimands could be useful in particular for theoretically motivated researchers who are seeking to test theories involving multiple treatments. Such theories not only predict whether one treatment should be more effective than another but also often dictate (a) the specific causal mechanisms that should grow or shrink when switching from one treatment to another and (b) the specific causal mechanisms that should contribute a larger share of the overall ATE for one treatment versus another. Indeed, for the purposes of theory testing and exploration, the two CCM estimands could be considered in conjunction with the ATEs to form a full picture of the relative causal anatomies of different treatments. To illustrate, Table 2 provides a set of some of the theoretical implications that would follow from testing hypotheses about the CCM estimands in combination with the ATEs.

Table 2: Theoretical Implications of Combined Hypotheses

| ACME_2 | ACME_1 | ACME_2/ATE_2 | ACME_1/ATE_1 |
|--------|--------|-------------|-------------|
| > 1    | > 1    | Disproportionate scaling up: Causal channel via M is larger in both absolute and proportional terms for second treatment. M is disproportionately responsible for enhancement of the effect when switching from first to second treatment. |
| yes    | yes    |            |            |
| no     | no     | Unrelatedness of mediator: The larger effect of the second treatment is not due to M. |
| yes    | no     | Proportionate scaling up: Causal channel via M is larger in absolute but not proportional terms for second treatment. M shares responsibility with other causal channels for enhancement of the effect when switching from first to second treatment. |
| no     | no     |            |            |

ATE_2 = ATE_1

yes yes Distinct causal anatomies: Despite equivalent ATEs, the treatments are comprised of differently sized causal channels, with M constituting a larger channel for the second treatment.

no no Indistinguishable causal anatomies: Any differences in the treatments’ causal anatomies are unrelated to M.

Note: Missing yes/no conditions are not applicable.
IV. COMPARATIVE CAUSAL MEDIATION ESTIMATION FRAMEWORK

Estimators of the CCM estimands, as presented in this study, will be formulated within a semi-parametric structural equations model (SEM) framework. A number of researchers have critiqued the parametric SEM framework for its inflexibility and reliance on functional form assumptions, instead advocating for more generalized, nonparametric formulations of causal mediation effects (Imai et al., 2010a, 2011b; Pearl, 2001, 2014). These are important and valid criticisms that researchers should keep in mind when considering the causal effects they are interested in exploring and estimating. However, the employment of a semi-parametric SEM framework is not necessarily restrictive in the context introduced by this study and can nest within the potential outcomes framework, as shown below.

A. Single Treatment Background

Before proceeding to the comparative causal mediation context, consider the single treatment causal mediation context. For a simple random sample of \( N \) independent observations, let \( Y_i(t, m) \) and \( M_i(t) \) denote the potential outcomes for subject \( i = 1, 2, ..., N \). Further, for any given subject, let the potential outcomes be fixed and characterized by the following:

\[
\begin{align*}
M_i(0) &= \pi_i \\
M_i(1) &= \pi_i + \alpha_i \\
Y_i(0, m) &= \lambda_i + \beta_i m \\
Y_i(1, m) &= (\lambda_i + \delta_i) + (\beta_i + \gamma_i)m
\end{align*}
\]

The relationships above implicitly assume that the potential outcomes are linear in \( m \), but provide for unit-specific parameters and hence are otherwise flexible given a binary treatment. In the case of a binary mediator in addition to a binary treatment, the relations would become fully flexible and non-parametric. The relations above can be equivalently expressed by the following equations:

\[
\begin{align*}
M_i &= \pi_i + \alpha_i T_i \\
Y_i &= \lambda_i + \delta_i T_i + \beta_i M_i + \gamma_i M_i T_i
\end{align*}
\]

These equations are written to share some notational similarities with the parametric SEMs often used to describe causal mediation, though a key difference is that the equations here allow for unit-specific parameters. As already described above, the only parametric constraint imposed given the binary treatment is that the outcome is linear with respect to the mediator, though in the case of a binary mediator, this no longer represents a constraint.

Under this semi-parametric set-up, we can highlight the relationship between the ACME as defined under the potential outcomes approach and the natural indirect effect as defined in the standard SEM approach to causal mediation:

\[^3\text{See }\text{Slipitser and VanderWeele (2011) and VanderWeele (2015) for a discussion of the connection between the nonparametric SEM and potential outcomes approaches to causal mediation analysis.}\]
\[ AMCE = \kappa(t) = E[Y_i(t, M_i(1)) - Y_i(t, M_i(0))] \]
\[ = E[\lambda_i + \delta_i t + \beta_i (\pi_i + \alpha_i) + \gamma_i t (\pi_i + \alpha_i)] - E[\lambda_i + \delta_i t + \beta_i (\pi_i) + \gamma_i t (\pi_i)] \]
\[ = E[\alpha_i (\beta_i + \gamma_i t)] \]

In the classic SEM framework (Baron and Kenny [1986], constant effects and no interaction between treatment and mediator are assumed under the structural equations \( M_i = \pi + \alpha T_i + \epsilon_i \) and \( Y_i = \lambda + \delta T_i + \beta M_i + \epsilon_i \). Applying those assumptions here, \( E[\alpha_i (\beta_i + \gamma_i t)] = \alpha \beta \), which is indeed the classic product-of-coefficients result in the SEM framework. However, this study will not assume constant effects, and the no-interaction assumption will be introduced but then relaxed.

**B. Two Treatment Set-Up**

Given two mutually exclusive treatment conditions (in addition to a control condition), the semi-parametric framework presented above can be easily extended. Now, for a simple random sample of \( N \) independent observations, let \( Y_i(t_1, t_2, m) \) and \( M_i(t_1, t_2) \) denote the potential outcomes for subject \( i = 1, 2, ..., N \). Given the mutual exclusivity of the two treatments, the following potential outcomes are defined:

\[
\begin{align*}
M_i(0, 0) & = \pi_i \\
M_i(1, 0) & = \pi_i + \alpha_{1i} \\
M_i(0, 1) & = \pi_i + \alpha_{2i} \\
Y_i(0, 0, m) & = \lambda_i + \beta_i m \\
Y_i(1, 0, m) & = (\lambda_i + \delta_{1i}) + (\beta_i + \gamma_{1i}) m \\
Y_i(0, 1, m) & = (\lambda_i + \delta_{2i}) + (\beta_i + \gamma_{2i}) m
\end{align*}
\]

As before, the relationships above assume that the potential outcomes are linear in \( m \), but are otherwise flexible given mutually exclusive, binary treatments. In the case of a binary mediator, the relations again would become fully flexible and non-parametric. The relations above yield:

\[
\begin{align*}
M_i & = \pi_i + \alpha_{1i} T_i + \alpha_{2i} T_{2i} \\
Y_i & = \lambda_i + \delta_{1i} T_i + \delta_{2i} T_{2i} + \beta_i M_i + \gamma_{1i} T_i M_i + \gamma_{2i} T_{2i} M_i \\
ACME_1 = \kappa_1(t_1) & = E[Y_i(t_1, 0, M_i(1, 0)) - Y_i(t_1, 0, M_i(0, 0))] = E[\alpha_{1i} (\beta_i + \gamma_{1i} t_1)] \\
ACME_2 = \kappa_2(t_2) & = E[Y_i(0, t_2, M_i(0, 1)) - Y_i(0, t_2, M_i(0, 0))] = E[\alpha_{2i} (\beta_i + \gamma_{2i} t_2)]
\end{align*}
\]

4The equivalency of the product of coefficients to the natural indirect effect is specific to the linear SEM formulation, though it has also been shown elsewhere to be a special case that nests within more general frameworks of causal mediation (Ji [2008] Pearl [2014]). This includes the potential outcomes framework, where it has been shown that the ACME is equivalent to \( \alpha \beta \) under certain conditions (Imai et al. [2010b]).
For each parameter $\theta_i$, define $\theta = E[\theta_i]$ and $\tilde{\theta}_i = \theta_i - \theta$. The equations above can thus be rewritten as follows:

$$M_i = \pi + \alpha_1 T_{1i} + \alpha_2 T_{2i} + \eta_i$$
$$Y_i = \lambda + \delta_1 T_{1i} + \delta_2 T_{2i} + \beta M_i + \gamma_1 T_{1i} M_i + \gamma_2 T_{2i} M_i + \iota_i$$

where

$$\eta_i = \tilde{\eta}_i + \delta_{1i} T_{1i} + \delta_{2i} T_{2i}$$
$$\iota_i = \tilde{\iota}_i + \tilde{\delta}_{1i} T_{1i} + \tilde{\delta}_{2i} T_{2i} + \tilde{\beta}_i M_i + \tilde{\gamma}_{1i} T_{1i} M_i + \tilde{\gamma}_{2i} T_{2i} M_i$$

In addition, also define the following:

$$Y_i = \chi + \tau_1 T_{1i} + \tau_2 T_{2i} + \rho_i$$
$$\rho_i = \tilde{\rho}_i + \tilde{\tau}_{1i} T_{1i} + \tilde{\tau}_{2i} T_{2i}$$
$$ATE_1 = \tau_1 = E[Y_i(1, 0, M_i(1, 0)) - Y_i(0, 0, M_i(0, 0))] = E[\tau_{1i}]$$
$$ATE_2 = \tau_2 = E[Y_i(0, 1, M_i(0, 1)) - Y_i(0, 0, M_i(0, 0))] = E[\tau_{2i}]$$

This yields the following set of semi-parametric equations where the individual-level heterogeneity is subsumed into the error terms:

$$M_i = \pi + \alpha_1 T_{1i} + \alpha_2 T_{2i} + \eta_i$$
$$Y_i = \lambda + \delta_1 T_{1i} + \delta_2 T_{2i} + \beta M_i + \gamma_1 T_{1i} M_i + \gamma_2 T_{2i} M_i + \iota_i$$

(5)

$$Y_i = \chi + \tau_1 T_{1i} + \tau_2 T_{2i} + \rho_i$$

(6)

$$Y_i = \chi + \tau_1 T_{1i} + \tau_2 T_{2i} + \rho_i$$

(7)

C. Assumptions

Recall that the estimands of interest are $\frac{\kappa_2(t_2)}{\kappa_1(t_1)}$ and $\left(\frac{\kappa_2(t_2)}{\kappa_1(t_1)}\right) / \left(\frac{\tau_1}{\tau_2}\right)$. The necessary assumptions for their consistent estimation are as follows.

The first identification assumption, which has already been implicit in the potential outcomes notation used up to this point, is the stable unit treatment value assumption (SUTVA).

**Assumption 1: Stable unit treatment value assumption (SUTVA)**

If $T_{1i} = T'_{1i}$, $T_{2i} = T'_{2i}$ and $M_i = M'_i$, then $Y_i(T_1, T_2, M) = Y_i(T'_1, T'_2, M')$ and $M_i(T_1, T_2) = Y_i(T'_1, T'_2)$, where $T_1$, $T_2$, and $M$ denote the full treatment and mediator vectors across subjects $i = 1, 2, \ldots, N$.

To be explicit, the linearity assumption is also reiterated.

**Assumption 2: Linear relationships between the potential outcomes and the mediator.**

$$Y_i(0, 0, m) = \lambda_i + \beta_i m$$
$$Y_i(1, 0, m) = (\lambda_i + \delta_{1i}) + (\beta_i + \gamma_{1i}) m$$
$$Y_i(0, 1, m) = (\lambda_i + \delta_{2i}) + (\beta_i + \gamma_{2i}) m$$

10
As already described above, while the assumption of linearity may seem rigid and demanding, this assumption is made trivial by the employment of a binary mediator. Given a binary mediator and the two mutually exclusive binary treatments, the potential outcome models described above are fully saturated and hence “inherently linear” [Angrist and Pischke 2009a, p. 37]. This is why it need not be stated nor assumed that the potential values of the mediator are linear in the treatments. This also helps to justify the exclusion of covariates from the model. In contrast to the case of estimating a single causal mediation effect, the CCM estimands can be estimated consistently without covariate adjustment, as will be shown shortly; furthermore, inclusion of covariates would invalidate the full saturation, and hence linearity, of the models.

The next assumption is that the two treatments, in addition to being mutually exclusive, have been completely randomized:

**Assumption 3:** Complete randomization of mutually exclusive treatments.

\[ P(T_1 = a, T_2 = b) = P(T_1 = a', T_2 = b') \]

for all \( a, a', b, \) and \( b' \) such that \( a^T b = a'^T b' = 0, 1^T a = 1^T a', \) and \( 1^T b = 1^T b' \) where \( 1 \) is the \( N \)-dimensional column vector with all elements equal to one.

The third assumption is that there is no interaction between the treatments and mediator in expectation.

**Assumption 4:** No expected interaction between the treatments and mediator.

\[ \gamma_1 = \gamma_2 = 0 \]

In other words, this assumption means that equation (6) becomes \( Y_i = \lambda + \delta_1 T_{1i} + \delta_2 T_{2i} + \beta M_i + \nu_i \). Compared to assumptions 2 and 3 which can be guaranteed by design, this assumption is somewhat stringent. However, this assumption will be relaxed \( (\gamma_1 \) and \( \gamma_2 \) will be allowed to be non-zero) later.

Finally, the last assumption pertains to the covariances between the individual-level parameters.

**Assumption 5:** No covariance between individual-level treatment and mediator parameters.

\[ \text{Cov}(\alpha_{1i}, \beta_i) = \text{Cov}(\alpha_{1i}, \gamma_{1i}) = 0 \]

\[ \text{Cov}(\alpha_{2i}, \beta_i) = \text{Cov}(\alpha_{2i}, \gamma_{2i}) = 0 \]

This type of no-covariance assumption is also implicitly made in other approaches to causal mediation. In the classic SEM formulation, the parameters are assumed to be constant structural effects, thereby meaning they do not vary across units and guaranteeing zero covariance across units. In addition, in the potential outcomes approach to causal mediation as applied
to a linear structural form, a conditional version of this assumption is implied by sequential ignorability. It is worth noting that a conditional version of this assumption is not necessarily any weaker or more plausible than an unconditional version, as there is no empirical or theoretical basis for expecting that any existing covariance between $\alpha_{ji}$ and $\beta_i$ will be attenuated within conditioning strata of the population. This is in contrast to omitted variable bias, which should generally be expected to shrink with stratification.

D. Consistent Estimation

Notably, the method presented here dispenses with the assumption of no confounding of the relationship between the mediator and outcome, which is a strong and nonrefutable assumption that is the most often criticized component of causal mediation analysis (e.g. Gerber and Green, 2012; Bullock et al., 2010; Glynn, 2012; Bullock and Ha, 2011). This assumption is required regardless of the statistical framework used for the identification and estimation of causal mediation effects, though its formal basis takes different forms depending on the statistical framework. In the SEM approach, this takes the form of recursivity or no correlation between the errors of the different equations, while in the potential outcomes framework, the unconfoundedness of the mediator-outcome relationship is implied by the “sequential ignorability” assumption. Notably, methods of sensitivity analysis have been developed to systematically assess the impact of violations of this assumption (e.g. Imai et al., 2010b). However, while such analyses allow for evaluation of the sensitivity of causal mediation estimates, they do not enable the recovery of consistent or unbiased estimates.

In the formulation here, such an assumption would take the form of $E[\iota_i|T_{1i}, T_{2i}, M_i] = 0$. Because the mediator has not been randomized, however, this assumption is difficult to justify and impossible to test; hence, this assumption will not be made. With the assumptions that are made, described above, it can be shown that estimation of $\beta$ via linear least squares regression results in the bias term $E[\hat{\beta} - \beta] = \frac{\text{cov}(\eta_i, \iota_i)}{\text{var}(\eta_i)}$. In contrast, $\alpha_j$ can be estimated consistently and without bias for both $j = 1, 2$. The key implication of these results is that, if comparing two treatments and their mediated effects via the same mediator, then a common bias afflicts both estimated ACMEs. By corollary, this means that the unavoidable mediation bias does not prevent us from comparing the causal mediation anatomies of two different treatments, as long as we are doing so in terms of the same mediator, thereby enabling us to gain important comparative insights on the mediation processes entailed by sets of treatments.

Proposition 1: Call $\hat{\tau}_2^N$, $\hat{\tau}_1^N$, $\hat{\alpha}_2^N$, $\hat{\alpha}_1^N$, and $\hat{\beta}^N$ the linear least squares regression estimators of the parameters from equations (5), (6), and (7) given a simple random sample of size $N$.

---

As Imai et al. (2010b) note, the sequential ignorability assumption implies a set of assumptions developed by Pearl (2001), which includes the independence between the potential values of the outcome and the potential values of the mediator. In the linear structural form, $\alpha_i$ is a function of the potential values of the mediator, while $\beta_i$ is a function of the potential values of the outcome. The independence between the potential values of the outcome and the potential values of the mediator implies the independence between these functions, thus implying independence between $\alpha_i$ and $\beta_i$. 

12
from a larger population. Given assumptions \(1, \ldots, 5\), then the following estimators converge in probability to the estimands of interest under the usual generalized linear regression regularity conditions:

\[
\text{plim}_{N \to \infty} \left( \frac{\hat{\alpha}_2^N \hat{\beta}_2^N}{\hat{\alpha}_1^N \hat{\beta}_1^N} \right) = \frac{\kappa_2(t_2)}{\kappa_1(t_1)} \quad \text{and} \quad \text{plim}_{N \to \infty} \left( \frac{\hat{\alpha}_1^N \hat{\beta}_1^N \hat{\tau}_2^N}{\hat{\alpha}_2^N \hat{\beta}_2^N \hat{\tau}_1^N} \right) = \frac{\kappa_2(t_2)}{\kappa_1(t_1)}
\]

In sum, the CCM estimands can be estimated consistently through the simple use of linear least squares regression estimators.

E. Uncertainty Estimation

Because the estimands of interest are nonlinear functions of multiple quantities that must each be estimated, the distributional behavior of the CCM estimators in finite samples is not immediately obvious. Furthermore, because the estimators employ ratios in which the distribution of the denominator may have positive probability density at zero, these estimators do not necessarily have finite-sample moments. This pathological problem is characteristic of ratio estimators in general, and it theoretically complicates the calculation of confidence intervals for those estimators. The existence of probability density at the point where the denominator equals zero creates a singularity in the distribution of a ratio estimator, which can result in the mysterious unbounded confidence interval. Yet traditional methods for constructing confidence sets do not necessarily take this property into account, and it has been shown that “any method which cannot generate unbounded confidence limits for a ratio leads to arbitrary large deviations from the intended confidence level” (von Luxburg and Franz, 2009; Gleser and Hwang, 1987; Koschat, 1987; Hwang, 1995). This issue has been studied extensively, with exact solutions having long ago been derived in special cases (e.g. Fieller, 1954) and approximation techniques based on the bootstrap developed recently for more general cases (Hwang, 1995; von Luxburg and Franz, 2009).

However, it has also been shown that in spite of the mathematical problems with ratio estimators, the use of standard methods for the practical estimation of confidence intervals yields approximately correct coverage under the reasonable condition that the confidence interval is actually bounded at the desired \(\alpha\) level, which is met when the \(1 - \alpha\) confidence interval of the denominator does not contain zero (Franz, 2007). These are also the conditions under which the CCM estimators can provide meaningful insights from a conceptual perspective. In those cases where the distribution of the estimated quantity in the denominator is not sufficiently bounded away from zero (i.e. where one of the mediation effects or proportions mediated is approximately zero), then it is not a conceptually meaningful quantity for use in these CCM

---

6Proofs of propositions can be found in Appendix A.

7And, as in any situation, a sufficiently large sample size is also necessary for analytic methods that rely on the central limit theorem and for bootstrap methods, which rely on the sample distribution adequately approximating the population distribution.
estimators. In other words, it falls outside of the meaningful scope of comparative causal mediation.

In sum, the potentially pathological nature of the CCM estimators is more of a mathematical inconvenience than a practical problem, just as is the case with the ratio-based Wald instrumental-variables estimator and just-identified 2SLS estimator more broadly. As long as the denominator is sufficiently bounded away from zero—i.e. significantly different from zero at the chosen \( \alpha \) level—then standard methods of confidence interval construction, including the Delta Method and standard bootstrap techniques, can be used as a practical matter. Furthermore, if the denominator is not sufficiently bounded away from zero, the variance and confidence intervals of the estimators would explode, similar to the behavior of the just-identified 2SLS estimator, protecting the researcher from drawing misguided conclusions.

Because the CCM estimands are only meaningful in the case that the denominator (for either estimand) is non-zero, it is recommended that a denominator that is statistically significantly different from zero at the chosen \( \alpha \) level be treated as a precondition for proceeding with CCM estimation. Not only is this recommended as a conceptual matter, but it also allows the researcher to avoid the problems of unbounded confidence intervals and to use the Delta Method and/or standard bootstrap techniques to compute confidence intervals.

F. Finite-Sample Adjustments

Due to their ratio form, the CCM estimators are not exactly centered on the true estimand in finite samples. As the sample size grows, this divergence becomes negligible. In smaller samples, finite-sample adjustments can be derived using Taylor series expansion. Appendix B presents the resulting adjusted estimators for both CCM estimands. Simulations, presented below, illustrate the improvement of the adjusted estimators over the simple estimators in small samples.

V. Simulations

To illustrate the properties of the CCM method, this section presents a simulation. Simulated causal mediation data were generated according to the following model, with the output of the first equation, \( M \), feeding into the second equation:

\[
M_i = \pi_i + \alpha_{1i}T_{1i} + \alpha_{2i}T_{2i} + \psi_iX_i
\]

\[
Y_i = \lambda_i + \delta_{1i}T_{1i} + \delta_{2i}T_{2i} + \beta_iM_i + \phi_iX_i
\]

\( T_1 \) and \( T_2 \) are indicator variable vectors and were generated such that an equal number of units were randomly assigned to (a) neither treatment, (b) \( T_1 \), and (c) \( T_2 \), with no units assigned to

---

8The analog in the instrumental variables estimation is a weak first-stage effect.

9For discussion of the finite-sample properties of just-identified 2SLS, including its lack of finite-sample moments and its unstable variance behavior given a weak first stage, see Angrist and Pischke (2009a, pp. 205-218), Angrist and Pischke (2009b), and Nelson and Starz (1988).
both $T_1$ and $T_2$. The rest of the variables and parameters were generated as follows:

$$X \sim Unif(0, 5) \quad \alpha_1 \sim N(4, 2) \quad \alpha_2 \sim N(10, 2) \quad \beta \sim N(3, 2)$$

$$\delta_1 \sim N(5, 2) \quad \delta_2 \sim N(5, 2) \quad \psi \sim N(4, 2) \quad \phi \sim N(4, 2) \quad \pi \sim N(0, 1) \quad \lambda \sim N(0, 1)$$

As indicated, the parameters were generated to vary independently across units, yielding heterogeneous effects with zero covariance between $\alpha_j$ and $\beta$ for $j = 1, 2$. Further, the data were also generated with no interaction between $T_j$ and $M$ for $j = 1, 2$. Along with the linear form and the exogeneity of $T_j$ for $j = 1, 2$, all assumptions established above are met by the data-generating process.

Once the data were generated, the mean values of the parameters $\alpha_1$, $\alpha_2$, and $\beta$—as well as $\tau_1$ and $\tau_2$—were estimated by linear least squares regression according to equations (5)-(7). Thus $X$ was omitted from the estimation process, simulating unobserved confounders and hence biasing the estimation of $\beta$ and the ACMEs.

In the results presented in Figure 1, the model was simulated 100 times with a total of 300 units per simulation (100 assigned to each of the two treatments and 100 assigned to neither treatment). Each panel in the plot displays the point estimates from each simulation for a different estimand, along with 95% confidence intervals constructed via the nonparametric percentile bootstrap. The panels in the top row correspond to the traditional causal mediation estimands: $ACME_1 (E[\alpha_1 \beta_i])$, $ACME_2 (E[\alpha_2 \beta_i])$, proportion of $ATE_1$ mediated ($\frac{E[\alpha_1 \beta_i]}{E[\tau_1]}$), and proportion of $ATE_2$ mediated ($\frac{E[\alpha_2 \beta_i]}{E[\tau_2]}$). The panels in the bottom row correspond to the CCM estimands, with both simple and small-sample adjusted estimators presented. The panels note the coverage of the confidence intervals, the true value of the estimand according to the data-generating process, and the mean estimate over all 100 simulations.

As can be seen, Figure 1 clearly shows how the traditional ACME estimators (top row) are biased and exhibit confidence-interval under-coverage given the presence of unmeasured confounders ($X$). The top left two panels show that the estimators of $ACME_1$ and $ACME_2$ are biased upward by approximately 2.5 and 6, resulting in only 90% and 72% coverage of the 95% confidence intervals. The story is the same for the top right two panels, which show the estimates of the proportions mediated for each treatment.

In contrast to the clear bias of the traditional causal mediation estimators, the bottom row shows that the CCM estimators are properly centered and exhibit good coverage. The bottom left two panels present the estimators of the ACME ratio, the first being the simple estimator and the second being the small-sample adjusted estimator. As can be seen, both perform well in recovering a mean estimate close to the true estimand value and proper confidence interval coverage (subject to simulation error). In addition, the small-sample adjustment also appears to yield a better mean estimate, demonstrating the gains from this adjustment. The results are the same in the bottom right two panels, which show the simple and adjusted estimators for the ratio of proportions mediated.
VI. RELAXING THE NO-INTERACTION ASSUMPTION

A. Set-Up

As before, recall the semi-parametric model:

\[
M_i = \pi + \alpha_1 T_{i1} + \alpha_2 T_{i2} + \eta_i
\]

\[
Y_i = \lambda + \delta_1 T_{i1} + \delta_2 T_{i2} + \beta M_i + \gamma_1 T_{i1} M_i + \gamma_2 T_{i2} M_i + \iota_i
\]

\[
Y_i = \chi + \tau_1 T_{i1} + \tau_2 T_{i2} + \rho_i
\]

The following will proceed without assumption [4] thereby allowing for treatment-mediator interactions, which has been referred to by some scholars as a version of moderated mediation [James and Brett, 1984; Preacher, 2007]. In this case, of interest are functions of the ACMEs for subsamples, namely for the treated units, \(\kappa_j(1)\), and for the control units, \(\kappa_j(0)\):

\[
\kappa_1(1) = E[\alpha_{1i}(\beta_i + \gamma_{1i})] = E[\alpha_{1i}\omega_{1i}] \quad \text{and} \quad \kappa_1(0) = E[\alpha_{1i}\beta_i]
\]

\[
\kappa_2(1) = E[\alpha_{2i}(\beta_i + \gamma_{2i})] = E[\alpha_{2i}\omega_{2i}] \quad \text{and} \quad \kappa_2(0) = E[\alpha_{2i}\beta_i]
\]

The same results as presented above (assuming no interactions) continue to apply in this case with regards to the ACMEs for the control units, \(\kappa_1(0)\) and \(\kappa_2(0)\). However, the CCM estimands are likely to be of greater theoretical and practical interest in terms of the ACMEs for the treated units. In this case, the estimands of interest are as follows:
Estimand 1: \[ \frac{\kappa_2(1)}{\kappa_1(1)} = \frac{E[\alpha_2 \omega_2]}{E[\alpha_1 \omega_1]} \]

Estimand 2: \[ \left( \frac{\kappa_2(1)}{\tau_2} \right) = \left( \frac{E[\alpha_2 \omega_2]}{E[\tau_2]} \right) \]

B. Conservatism of Estimators

Call \( \hat{\tau}_2, \hat{\tau}_1, \hat{\alpha}_2, \hat{\alpha}_1, \hat{\beta}, \hat{\gamma}_2, \) and \( \hat{\gamma}_1 \) the linear least squares regression estimators of the parameters from equations (5), (6), and (7). Once again, the randomization of the treatments guarantees consistency for \( \hat{\tau}_2, \hat{\tau}_1, \) and \( \hat{\alpha}_1 \) under standard regularity conditions, but not for \( \hat{\beta}, \hat{\gamma}_2, \) and \( \hat{\gamma}_1 \). Under certain conditions, it can be shown that \( \frac{\hat{\alpha}_2(\hat{\beta} + \hat{\gamma}_2)}{\hat{\alpha}_1(\hat{\beta} + \hat{\gamma}_1)} \) and \( \frac{\hat{\alpha}_2(\hat{\beta} + \hat{\gamma}_2)}{\hat{\alpha}_1(\hat{\beta} + \hat{\gamma}_1)} \) are not consistent estimators of \( \frac{\kappa_2(1)}{\kappa_1(1)} \) and \( \frac{\kappa_2(1)}{\tau_2} \), respectively, but are asymptotically conservative (attenuated toward unity). These simple estimators are conservative only in the probability limit because, as before, there is a finite-sample divergence due to the ratio form of the estimators. However, also as before, that finite-sample divergence can be approximated, estimated, and used to construct adjusted estimators.

**Proposition 2:** Without loss of generality, assume that both the numerator and denominator of the estimator are positive, and that the estimator is greater than 1 (i.e. the numerator is larger than the denominator). Call \( \hat{\tau}_2^N, \hat{\tau}_1^N, \hat{\alpha}_2^N, \hat{\alpha}_1^N, \hat{\beta}_N, \hat{\gamma}_2^N, \hat{\gamma}_1^N \) the linear least squares regression estimators of the parameters from equations (5), (6), and (7) given a simple random sample of size \( N \) from a larger population. Let \( \hat{\omega}_1^N = \hat{\beta}_N + \hat{\gamma}_1^N \) and \( \hat{\omega}_2^N = \hat{\beta}_N + \hat{\gamma}_2^N \). Further call \( \xi_1 \) and \( \xi_2 \) the asymptotic bias components of \( \hat{\omega}_1^N \) and \( \hat{\omega}_2^N \), respectively (i.e. \( \text{plim}_{N \to \infty} \hat{\omega}_1^N - \omega_1 = \xi_1 \) and \( \text{plim}_{N \to \infty} \hat{\omega}_2^N - \omega_2 = \xi_2 \)). Make assumptions [4], [2], [4], and [2]. Then, given \( \omega_2 \xi_1 > \omega_1 \xi_2 \), the following holds:

\[
\text{plim}_{N \to \infty} \frac{\hat{\alpha}_2^N \omega_2^N}{\hat{\alpha}_1^N \omega_1^N} < \frac{\kappa_2(1)}{\kappa_1(1)}
\]

\[
\text{plim}_{N \to \infty} \frac{\hat{\alpha}_2^N \omega_2^N}{\hat{\alpha}_1^N \omega_1^N} \left( \frac{\hat{\alpha}_2^N \omega_2^N}{\hat{\alpha}_1^N \omega_1^N} \right) < \left( \frac{\kappa_2(1)}{\tau_2} \right)
\]

The result is that, given the conditions described in Proposition 2, the bias attenuates the estimates of the two CCM estimands. Since these results were presented without loss of generality in the context where the estimands are greater than 1, this means that the attenuated estimates will be conservative. In other words, the estimates will be biased in favor of the null hypothesis that the estimands equal 1.

While assumption [4] (no interaction between the treatments and mediator) was relaxed, Proposition 2 introduces the following additional condition that was not present in Proposition [4] \( \omega_2 \xi_1 > \omega_1 \xi_2 \). As shown in Appendix C, this condition can be partially assessed empirically.

[Loeys et al. 2016] describe specific conditions under which \( \hat{\gamma}_2 \) and \( \hat{\gamma}_1 \) are unbiased estimators even when \( \hat{\beta} \) is not.
C. Additional Notes

Similar to the case in which the no-interaction assumption is maintained, finite-sample adjustments can be derived for the CCM estimators when relaxing the no-interaction assumption. Appendix B presents these finite-sample adjustments. In addition, Appendix D presents simulation results to illustrate the properties of the CCM estimators when the no-interaction assumption has been relaxed.

VII. Application: International Law and Audience Costs

A. Background

Does international law affect state behavior? There is a longstanding scholarly debate on this question, with some political scientists and legal scholars viewing international law as largely epiphenomenal to state interests and power (e.g. Downs et al. 1996; Goldsmith and Posner 2005), and others seeing international law as having a real impact on state decision-making (e.g. Goldstein 2001). Among the latter group, many scholars have identified domestic political processes and institutions as an important conduit through which national governments can be induced to honor their international legal obligations, even in cases where those governments did not intend to comply in the first place (Simmons 2009; Trachtman 2010; Hathaway 2002; Moravcsik 2013; Dai 2005; Abbott and Snidal 1998; Risse-Kappen et al. 1999). The electoral compliance mechanism, in which governments are incentivized to maintain compliance with international legal agreements under the threat of electoral punishment for violations, is one possible domestic source of compliance.

In a number of recent studies using survey experiments, political scientists have accumulated evidence that voters in the United States and elsewhere are indeed inclined to punish elected officials who renege on previous foreign policy commitments (Tomz 2007; McGillivray and Smith 2000; Chaudoin 2014; Chilton 2015; Hyde 2015). The political costs that a government incurs as a result of constituents disapproving of violations of policy commitments—which may manifest in the form of electoral power in democracies or via the threat of protest and dissent in non-democracies—are generally referred to as domestic “audience costs” (Fearon 1994; Morrow 2000; Tomz 2007; Weeks 2008; Jensen 2003).

However, an important gap remains in this scholarship: while studies have shown that public disapproval of a foreign policy decision tends to increase when that policy decision requires reneging on international legal commitments, these studies have not isolated the role of legality per se in generating that disapproval. Instead, the design of these studies has masked the extent to which such disapproval is attributable to the baseline breaking of the commitment (i.e. the audience costs for not honoring a policy pledge in general) versus the additional legal status of the commitment. In other words, we do not know whether the dimension of international legality actually enhances audience costs, and if it does, to what extent or why that is the case.

Indeed, in scholarship on public attitudes toward international commitments, much of the
international relations literature tends to abstract away the distinctive nature of legality and treat international legal commitments as generic international commitments. The implications of such a framing is that legality should not affect the prospect for audience costs. Yet there are, of course, reasons to believe that voters will respond more negatively to home government violations of foreign policy commitments when those violations also entail breaking international law. Voters may view legal commitments as uniquely serious and solemn forms of commitment, the violation of which is considered particularly objectionable, in which case legality should increase the prospect for audiences costs. While this has been suggested in the literature (Lipson, 1991; Abbott and Snidal, 2000a; Simmons and Hopkins, 2005), it has not been explicitly tested.

B. Study Design

In order to address this gap in the literature, the author designed and implemented a novel survey experiment embedded in an online survey administered in August 2015, with 1602 U.S.-based respondents recruited via Amazon Mechanical Turk. The experiment revolved around a security scenario in which the U.S. government decided to take military action against ISIS forces in Iraq.

The scenario involved a U.S. military operation in Iraq to capture ISIS militants who were threatening rocket attacks on neighboring countries but were hiding in a civilian zone. Respondents were told that in order to avoid collateral damage, the U.S. military deployed commandos in a covert operation, in which the commandos used an ostensibly non-lethal incapacitating chemical gas to neutralize the ISIS militants. The incapacitating gas was featured in the scenario in order to exploit real-world ambiguity surrounding the international legality of chemical incapacitants in unconventional operations, as well as ambiguity surrounding the lethality of these chemical agents. Because of this ambiguity and the technical nature of the legal categorization of chemical incapacitants, survey respondents should not be expected to identify such agents as clearly illegal, in contrast to well-known chemical warfare agents. At the same time, it is also plausible and hence reasonable to convince respondents that these chemical incapacitants are illegal under the Chemical Weapons Convention. As a result, it was possible to effectively intervene upon respondents’ knowledge of the legal status of these chemical incapacitants. The survey instrument text and variable coding rules can be found in Appendix E, and sample demographic distributions can be found in Appendix F.

There were two primary goals of the research. The first goal was to disentangle the dimension of (il)legality from the baseline violation of a foreign policy commitment more explicitly than have previous studies, thereby creating a more valid design to answer the research question: Does the international legal status of a foreign policy commitment increase the potential

\[11\] This research was approved by the Institutional Review Board at Stanford University (Protocol 31139).

\[12\] While the illegality of chemical incapacitants is probably the most widely accepted position among arms control legal experts, some experts have argued otherwise in terms of the use of chemical incapacitants under certain conditions. For an overview of the debate, see Ballard (2007).
for domestic audience costs if that commitment is violated? To achieve this goal, the experimental design featured two mutually exclusive treatment conditions in addition to a control condition. In the control condition, respondents were simply told about the U.S. government’s decision to use military force employing chemical incapacitants. In the first “informal” treatment condition, respondents were additionally told that this decision constituted a violation of the U.S. government’s previous foreign policy commitment, but they were not given any information about international legality. In the second “legal” treatment condition, respondents were told that this decision constituted a violation of the U.S. government’s international legal commitment. There were two binary outcome variables of interest. The first measured whether or not the respondent disapproved of the policy decision to use chemical incapacitants, which will be called Disapproval. The second measured whether or not the respondent would be less likely to vote for a U.S. Senator who supported the policy decision, which will be called Punishment.

The second research goal was to identify and better understand the contours of public opinion that determine the extent to which legalization does (or does not) amplify audience costs. In addition to measuring Disapproval and Punishment, other attitudes were also measured. The primary attitude of interest, under investigation as a mediator, was the respondents’ perception of the (im)morality of the decision to use chemical incapacitants. Perceptions of immorality represent a key theoretical reason, long noted by scholars, that voters would more strongly disapprove of violations of legalized foreign policy commitments than similar non-legalized commitments (e.g. Abbott and Snidal 2000b, pp. 428-429). Hence, a binary mediator variable measured whether or not each respondent believed the policy decision to be immoral, which will be called Perceived Immorality. This enables estimation of the portion of each treatment effect, $ATE_1$ (informal) and $ATE_2$ (legal), that is transmitted via Perceived Immorality—that is, estimation of $ACME_1$ and $ACME_2$.

As described above, the problem with traditional mediation analysis is that, even with pre-treatment covariates included as controls, those mediated effects are likely to be biased and inconsistent. However, under the assumptions stated earlier, the CCM estimands can be estimated consistently (or conservatively). The first estimand $\frac{ACME_2}{ACME_1}$ measures the extent to which the morality mediator transmits a stronger effect for the legal treatment than for the informal treatment. The second estimand $(\frac{ACME_2}{ATE_2})/(\frac{ACME_1}{ATE_1})$ measures the extent to which the morality mediator comprises a larger proportion of the total effect of (i.e. is more important for) the legal treatment, compared to the informal treatment.

C. Results

The results of the survey experiment provide statistically and substantively strong evidence that the legal treatment does indeed cause a larger increase in the probability of Disapproval and Punishment than the informal treatment, as shown by Table 3, providing support for the theory that legalization enhances audience costs. Specifically, the legal treatment had
an estimated 12.5 percentage-point larger effect on the probability of Disapproval and a 9.9 percentage-point larger effect on the probability of Punishment than the informal treatment.

Table 3: Sample Estimates of ATEs

| DV: Disapproval |  |  |
|-----------------|-----------------|-----------------|
| $\hat{ATE}_1$   | $\hat{ATE}_2$   | $\hat{ATE}_2 - \hat{ATE}_1$ |
| (Informal treatment effect) | (Legal treatment effect) | (Difference in treatment effects) |
| Estimate | 0.195 | 0.320 | 0.125 |
| Bootstrap 95% CI | [0.134, 0.249] | [0.265, 0.372] | [0.065, 0.181] |

| DV: Punishment |  |  |
|-----------------|-----------------|-----------------|
| $\hat{ATE}_1$   | $\hat{ATE}_2$   | $\hat{ATE}_2 - \hat{ATE}_1$ |
| (Informal treatment effect) | (Legal treatment effect) | (Difference in treatment effects) |
| Estimate | 0.182 | 0.281 | 0.099 |
| Bootstrap 95% CI | [0.127, 0.234] | [0.226, 0.333] | [0.039, 0.156] |

More importantly in the context of this study, however, the results of the CCM analysis also provided support for the theory that this enhancement of audience costs by legalization is, at least in part, due to an increase in Perceived Immorality. Table 4 shows the results of the CCM analysis. The assumption of no interaction between the treatments and mediator was tested in the case of both dependent variables. The test failed to reject the null hypothesis of no interactions in the case of the Disapproval dependent variable, and hence the no-interaction assumption was maintained in that case.

However, the test rejected the null hypothesis of no interactions in the case of the Punishment dependent variable, which is why the causal mediation estimates in the Punishment case involve the ACMEs for the treated (ACMETs)—that is $\kappa_1(1)$ and $\kappa_2(1)$. Furthermore, additional tests provide support for the conditions necessary for the CCM estimators to be conservative given the interactions between the treatments and mediator. Specifically, the tests provide evidence that $\omega_2 \xi_1 > \omega_1 \xi_2$.\(^{13}\)

The CCM estimates are presented in bold in Table 4. Given the large sample size, these estimates were obtained using the simple estimators\(^{14}\) and the 95% confidence intervals were computed via the nonparametric percentile bootstrap. As can be seen, the $\frac{\text{ACME}_2}{\text{ACME}_1}$ esti-

---

\(^{13}\)As explained in Appendix C, this is tested partially by verifying that $\hat{\omega_2 \text{Var}(M_i | T_1 = 0, T_2 = 1)} > \hat{\omega_1 \text{Var}(M_i | T_1 = 1, T_2 = 0)}$.

\(^{14}\)The finite-sample adjusted estimates are virtually identical, as should be expected given the sample size. For
ates are statistically (and substantively) distinguishable from 1 for both dependent variables. These estimates can be interpreted as meaning that the effect on Disapproval (Punishment) mediated via Perceived Immorality is about 56% (83%) larger for the legal treatment than for the informal treatment. In contrast, the \( \frac{\text{ACME}_2}{\text{ATE}_2} / \frac{\text{ACME}_1}{\text{ATE}_1} \) estimates are not statistically distinguishable from 1 for either dependent variable. This means that while Perceived Immorality transmitted a larger effect for the legal treatment than the informal treatment, it did not necessarily constitute a larger proportion of the overall ATE for the legal treatment compared to the informal treatment.

Table 4: Comparative Causal Mediation Analysis

| DV: Disapproval | ACME \_2 | ACME \_1 | ACME \_2 / ACME \_1 | ACME \_2 |
|----------------|---------|---------|-----------------|--------|
| Estimate       | 1.563   | 0.952   | 0.113           | 0.177  |
| Bootstrap 95% CI | [1.189, 2.170] | [0.752, 1.212] | [0.078, 0.149] | [0.140, 0.215] |

| DV: Punishment | ACMET \_2 | ACMET \_1 | ACMET \_2 / ACMET \_1 | ACMET \_2 |
|----------------|---------|---------|-----------------|--------|
| Estimate       | 1.829   | 1.184   | 0.096           | 0.176  |
| Bootstrap 95% CI | [1.321, 2.657] | [0.886, 1.635] | [0.064, 0.131] | [0.137, 0.213] |

In combination, the results correspond to the case of “proportionate scaling up” presented in Table 2. Perceived Immorality is found to be an important factor that leads to a scaling up of the audience costs effect given legalization. Yet it appears that other mediation channels also help to scale up that effect such that while the mediation channel via Perceived Immorality expands, it does not increase as a proportion of the total effect. Overall, the collective evidence provides support in favor of the theory that international legalization enhances audience costs (at least partially) by amplifying the perceived immorality of violating a foreign policy commitment.

In particular, the adjusted estimate of \( \frac{\text{ACME}_2}{\text{ACME}_1} \) for the Disapproval dependent variable is 1.533, and the adjusted estimate of \( \frac{\text{ACMET}_2}{\text{ACMET}_1} \) for the Punishment dependent variable is 1.796.
As explained in this study, it is difficult to have confidence in the accuracy of the individual ACME estimates, even if covariates were to be included as controls, because this would require a strong assumption of no unobserved confounding of the mediator-outcome relationship. However, under a weaker set of assumptions, one can have greater confidence in the estimates of the CCM estimands, which also happen to be the estimands of primary interest for the particular theory and hypotheses at stake in this and many other applications.

VIII. Conclusions

This study has introduced a novel set of causal mediation estimands which compare the causal mediation effects of multiple treatments. It has shown that these estimands can be estimated consistently or conservatively under weaker assumptions than can any single average causal mediation effect (ACME). In particular, the usual assumption of no confounding of the mediator-outcome relationship, which is required for consistent estimation of a single ACME, is not necessary in the comparative causal mediation context presented in this study.

With the gradual accumulation of knowledge and empirical results in various academic sub-fields and program evaluation contexts, experimental research increasingly involves evaluating multiple treatments—that is, investigating the relative strengths and comparing the causal anatomies of distinct but conceptually or administratively related treatments—rather than simply testing the effects of single treatments. The method of CCM analysis presented in this study provides a new tool for researchers who are interested in comparing, discovering, and testing the causal mechanism differences between multiple treatments, and would like to do so under the weakest possible set of assumptions.
References

Aaroe, L. (2012). When citizens go against elite directions: Partisan cues and contrast effects on citizens attitudes. *Party Politics, 18*(2):215–233.

Abbott, K. W. and Snidal, D. (1998). Why states act through formal international organizations. *Journal of conflict resolution, 42*(1):3–32.

Abbott, K. W. and Snidal, D. (2000a). Hard and soft law in international governance. *International organization, 54*(3):421–456.

Abbott, K. W. and Snidal, D. (2000b). Hard and soft law in international governance. *International Organization, 54*(3):421–456.

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in medicine, 27*(8):1282–1304.

Angrist, J. D. and Pischke, J.-S. (2009a). *Mostly Harmless Econometrics: An Empiricists Companion.* Princeton: Princeton University Press.

Angrist, J. D. and Pischke, J.-S. (2009b). A note on bias in just identified iv with weak instruments. [http://econ.lse.ac.uk/staff/spischke/mhe/josh/solon_justid_april14.pdf](http://econ.lse.ac.uk/staff/spischke/mhe/josh/solon_justid_april14.pdf). Technical report, J. Angrist and J.S. Pischke.

Arceneaux, K. and Kolodny, R. (2009). Educating the least informed: Group endorsements in a grassroots campaign. *American Journal of Political Science, 53*(4):755–770.

Ballard, K. (2007). Convention in peril? riot control agents and the chemical weapons ban. *Arms Control Today, 37*(7).

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.

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

Bullock, J. G. and Ha, S. E. (2011). Mediation analysis is harder than it looks. In Druckman, J. N., Green, D. P., Kuklinski, J. H., and Lupia, A., editors, *Cambridge Handbook of Experimental Political Science*, chapter 35, pages 508–521. Cambridge University Press.

Chaudoin, S. (2014). Promises or policies? an experimental analysis of international agreements and audience reactions. *International Organization, 68*(1):235–256.

Chilton, A. S. (2015). The laws of war and public opinion: An experimental study. *Journal of Institutional and Theoretical Economics JITE, 171*(1):181–201.

Dai, X. (2005). Why comply? the domestic constituency mechanism. *International Organization, 59*(2):363–398.
Daniel, R. M., DeStavola, B. L., Cousens, S. N., and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14.

Downs, G. W., Rocke, D. M., and Barsoom, P. N. (1996). Is the good news about compliance good news about cooperation? *International Organization*, 50(3):379–406.

Fearon, J. D. (1994). Domestic political audiences and the escalation of international disputes. *American Political Science Review*, 88(3):577–592.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B*, 16(2):175–185.

Franz, V. H. (2007). Ratios: A short guide to confidence limits and proper use (http://arxiv.org/abs/0710.2024). Technical report, arXiv.org.

Gerber, A. S. and Green, D. P. (2012). Mediation. In *Field Experiments: Design, Analysis, and Interpretation*, chapter 10. New York: W. W. Norton & Company.

Gleser, L. J. and Hwang, J. T. (1987). The nonexistence of 100(1-alpha)% confidence sets of finite expected diameter in errors-in-variables and related models. *The Annals of Statistics*, 15(4):1351–1362.

Glynn, A. N. (2012). The product and difference fallacies for indirect effects. *American Journal of Political Science*, 56(1):257–269.

Goldsmith, J. L. and Posner, E. A. (2005). *The limits of international law*. Oxford University Press.

Goldstein, J. (2001). *Legalization and world politics*. MIT Press.

Goren, P., Federico, C. M., and Kittilson, M. C. (2009). Source cues, partisan identities, and political value expression. *American Journal of Political Science*, 53(4):805–820.

Hathaway, O. A. (2002). Do human rights treaties make a difference? *The Yale Law Journal*, 111(8):1935–2042.

Hwang, J. T. G. (1995). Fieller’s problems and resampling techniques. *Statistica Sinica*, 5:161–171.

Hyde, S. D. (2015). Experiments in international relations: Lab, survey, and field. *Annual Review of Political Science*, 18:403–424.

Imai, K., Jo, B., and Stuart, E. A. (2011a). Commentary: Using potential outcomes to understand causal mediation analysis. *Multivariate Behavioral Research*, 46(5):842–854.

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. (2011b). 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. (2010b). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25:51–71.

Imai, K. and Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171.

James, L. R. and Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69(2):307–321.

Jensen, N. M. (2003). Democratic governance and multinational corporations: Political regimes and inflows of foreign direct investment. *International Organization*, 57(3):587–616.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, 13:314–336.

Koschat, M. A. (1987). A characterization of the fieller solution. *The Annals of Statistics*, 15(1):462–468.

Lipson, C. (1991). Why are some international agreements informal? *International Organization*, 45(4):495–538.

Loeys, T., Talloen, W., Goubert, L., Moerkerke, B., and Vansteelandt, S. (2016). Assessing moderated mediation in linear models requires fewer confounding assumptions than assessing mediation. *British Journal of Mathematical and Statistical Psychology*, 69(3):352–374.

McGillivray, F. and Smith, A. (2000). Trust and cooperation through agent-specific punishments. *International Organization*, 54(4):809–824.

Moravcsik, A. (2013). Liberal theories of international law. In Dunoff, J. L. and Pollack, M. A., editors, *Interdisciplinary Perspectives on International Law and International Relations*, chapter 4, pages 83–118. Cambridge University Press, Cambridge.

Morrow, J. D. (2000). Alliances: Why write them down? *Annual Review of Political Science*, 3(1):63–83.

Nelson, C. R. and Starz, R. (1988). Some further results on the exact small sample properties of the instrumental variable estimator. Technical report, NBER Technical Working Paper No. 68.

Nicholson, S. P. (2012). Polarizing cues. *American Journal of Political Science*, 56(1):52–66.

Pearl, J. (2001). Direct and indirect effects. Technical report, Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence.
Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, 19(4):459–481.

Preacher, K. J. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1):185–227.

Risse-Kappen, T., Ropp, S. C., and Sikkink, K. (1999). *The power of human rights: International norms and domestic change*, volume 66. Cambridge University Press.

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

Shpitser, I. and VanderWeele, T. J. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *The International Journal of Biostatistics*, 7(1).

Simmons, B. A. (2009). *Mobilizing for human rights: international law in domestic politics*. Cambridge University Press.

Simmons, B. A. and Hopkins, D. J. (2005). The constraining power of international treaties: Theory and methods. *American Political Science Review*, 99(4):623–631.

Slothuus, R. and de Vreese, C. H. (2010). Political parties, motivated reasoning, and issue framing effects. *American Journal of Political Science*, 72(3):630–645.

Tomz, M. (2007). Domestic audience costs in international relations: An experimental approach. *International Organization*, 61:821–840.

Trachtman, J. P. (2010). International law and domestic political coalitions: The grand theory of compliance with international law. *Chicago Journal of International Law*, 11:128–129.

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

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.

von Luxburg, U. and Franz, V. H. (2009). A geometric approach to confidence sets for ratios: Fieller’s theorem, generalizations and bootstrap. *Statistica Sinica*, 19:1095–1117.

Weeks, J. L. (2008). Autocratic audience costs: Regime type and signaling resolve. *International Organization*, 62(1):35–64.
Supplementary Materials

for

Comparative Causal Mediation and Relaxing the Assumption of No Mediator-Outcome Confounding: An Application to International Law and Audience Costs

Kirk Bansak
APPENDIX A: Formal Results

Proof of Proposition 1

Given assumptions 1 and 2

\[ \kappa_j(t_j) = E[\alpha_{ji}(\beta_i + \gamma_{ji}t_j)] = E[\alpha_{ji}\beta_i] + E[\alpha_{ji}\gamma_{ji}t_j] \]

for \( j = 1, 2 \).

Given assumption 5

\[ E[\alpha_{ji}\beta_i] + E[\alpha_{ji}\gamma_{ji}t_j] = E[\alpha_{ji}]E[\beta_i] + E[\alpha_{ji}]E[\gamma_{ji}]t_j = \alpha_j(\beta + \gamma t_j) \]

for \( j = 1, 2 \).

Given assumption 4

\[ \alpha_j(\beta + \gamma_{ji}t_j) = \alpha_j\beta \]

for \( j = 1, 2 \).

Thus,

\[ \frac{\kappa_2(t_2)}{\kappa_1(t_1)} = \frac{\alpha_2\beta}{\alpha_1\beta} \]

and

\[ \frac{\frac{\kappa_2(t_2)}{\tau_2}}{\frac{\kappa_1(t_1)}{\tau_1}} = \frac{\frac{\alpha_2\beta}{\tau_2}}{\frac{\alpha_1\beta}{\tau_1}} \]

Now, given assumption 3

\[ E[\eta_i|T_{1i}, T_{2i}] = E[\tilde{\pi}_i + \tilde{\alpha}_1iT_{1i} + \tilde{\alpha}_2iT_{2i}|T_{1i}, T_{2i}] \]
\[ = E[\tilde{\pi}_i|T_{1i}, T_{2i}] + E[\tilde{\alpha}_1iT_{1i}, T_{2i}]T_{1i} + E[\tilde{\alpha}_2iT_{1i}, T_{2i}]T_{2i} \]
\[ = E[\tilde{\pi}_i] + E[\tilde{\alpha}_1iT_{1i}] + E[\tilde{\alpha}_2iT_{2i}] \]
\[ = E[\pi_i - \pi] + E[\alpha_{1i} - \alpha_1]T_{1i} + E[\alpha_{2i} - \alpha_2]T_{2i} \]
\[ = 0 \]
and

\[
\begin{align*}
E[\rho_i|T_{1i}, T_{2i}] &= E[\tilde{\chi}_i + \tilde{\tau}_1 T_{1i} + \tilde{\tau}_2 T_{2i}|T_{1i}, T_{2i}] \\
&= E[\tilde{\chi}_i|T_{1i}, T_{2i}] + E[\tilde{\tau}_1|T_{1i}, T_{2i}]T_{1i} + E[\tilde{\tau}_2|T_{1i}, T_{2i}]T_{2i} \\
&= E[\tilde{\chi}_i] + E[\tilde{\tau}_1]T_{1i} + E[\tilde{\tau}_2]T_{2i} \\
&= E[\chi_i - \chi] + E[\tau_1 - \tau_1]|T_{1i} + E[\tau_2 - \tau_2]|T_{2i} \\
&= 0
\end{align*}
\]

Therefore, under standard regularity conditions for generalized linear regression model,

\[
\begin{align*}
\text{plim}_{N \to \infty} \hat{\alpha}_1^N &= \alpha_1 \\
\text{plim}_{N \to \infty} \hat{\alpha}_2^N &= \alpha_2 \\
\text{plim}_{N \to \infty} \hat{\tau}_1^N &= \tau_1 \\
\text{plim}_{N \to \infty} \hat{\tau}_2^N &= \tau_2
\end{align*}
\]

Further, by Slutsky’s theorem, and given non-zero parameters,

\[
\begin{align*}
\text{plim}_{N \to \infty} \left( \frac{\hat{\alpha}_2^N \hat{\beta}_2^N}{\hat{\alpha}_1^N \hat{\beta}_1^N} \right) = \text{plim}_{N \to \infty} \left( \frac{\hat{\alpha}_2^N}{\hat{\alpha}_1^N} \right) = \left( \text{plim}_{N \to \infty} \hat{\alpha}_2^N \right) / \left( \text{plim}_{N \to \infty} \hat{\alpha}_1^N \right) = \frac{\alpha_2}{\alpha_1} = \frac{\alpha_2 \beta}{\alpha_1 \beta} = \frac{\kappa_2}{\kappa_1}(t_2) \\
\end{align*}
\]

And by the same argument

\[
\begin{align*}
\text{plim}_{N \to \infty} \left( \frac{\hat{\alpha}_2^N \hat{\beta}_2^N}{\hat{\alpha}_1^N \hat{\beta}_1^N} \right) = \frac{\left( \frac{\alpha_2 \beta}{\tau_2} \right) \left( \frac{\alpha_1 \beta}{\tau_1} \right)}{\left( \frac{\kappa_2(t_2)}{\tau_2} \right) \left( \frac{\kappa_1(t_1)}{\tau_1} \right)}
\end{align*}
\]
Proof of Proposition 2

Given assumptions 1 and 2

\[ \kappa_j(1) = E[\alpha_{ji}(\beta_i + \gamma_{ji})] = E[\alpha_{ji}\omega_{ji}] \]

for \( j = 1, 2 \).

Given assumption 5

\[ E[\alpha_{ji}\omega_{ji}] = E[\alpha_{ji}]E[\omega_{ji}] = \alpha_{ji}\omega_j \]

for \( j = 1, 2 \).

Thus,

\[ \frac{\kappa_2(1)}{\kappa_1(1)} = \frac{\alpha_2\omega_2}{\alpha_1\omega_1} \]

and

\[ \frac{\kappa_2(1)}{\tau_2} \frac{\kappa_1(1)}{\tau_1} = \frac{\alpha_2\omega_2}{\tau_2} \frac{\alpha_1\omega_1}{\tau_1} \]

Now, given assumption 3 as in the proof of Proposition 1 under standard regularity conditions,

\[ \text{plim}_{N \to \infty} \hat{\alpha}_1^N = \alpha_1 \]

\[ \text{plim}_{N \to \infty} \hat{\alpha}_2^N = \alpha_2 \]

\[ \text{plim}_{N \to \infty} \hat{\tau}_1^N = \tau_1 \]

\[ \text{plim}_{N \to \infty} \hat{\tau}_2^N = \tau_2 \]

It will thus be the case that

\[ \text{plim}_{N \to \infty} \frac{\hat{\alpha}_2^N \hat{\omega}_2^N}{\hat{\alpha}_1^N \hat{\omega}_1^N} < \frac{\alpha_2\omega_2}{\alpha_1\omega_1} \]

and

\[ \text{plim}_{N \to \infty} \left( \frac{\hat{\alpha}_2^N \hat{\omega}_2^N}{\hat{\tau}_2^N} \right) \left( \frac{\hat{\alpha}_1^N \hat{\omega}_1^N}{\hat{\tau}_1^N} \right) < \frac{\alpha_2\omega_2}{\tau_2} \frac{\alpha_1\omega_1}{\tau_1} \]

if

\[ \text{plim}_{N \to \infty} \frac{\hat{\omega}_2^N}{\hat{\omega}_1^N} < \frac{\omega_2}{\omega_1} \]
which is met when: \[\frac{\omega_2 + \xi_2}{\omega_1 + \xi_1} < \frac{\omega_2}{\omega_1}\]

and hence when:

\[\omega_1 \xi_2 < \omega_2 \xi_1\]
Appendix B: Finite-Sample Adjustments

Finite-sample adjustments for the CCM estimators can be derived using Taylor series expansion.

Consider the first estimator under the no-interaction assumption, \( \frac{\hat{\alpha}_2 \hat{\beta}}{\hat{\alpha}_1 \hat{\beta}} \), which can (quite apparently) be simplified to \( \frac{\hat{\alpha}_2}{\hat{\alpha}_1} \). Similarly, the estimand of interest can be seen simply as:

\[
ACME_2 = \frac{\alpha_2 \beta}{\alpha_1 \beta} = \frac{\alpha_2}{\alpha_1} = E[\hat{\alpha}_2] \quad \frac{\hat{\alpha}_1}{E[\hat{\alpha}_1]}
\]

However, a first problem is the following:

\[
E[\hat{\alpha}_2 \hat{\alpha}_1] \neq E[\hat{\alpha}_2] \frac{\hat{\alpha}_1}{E[\hat{\alpha}_1]} = \frac{\alpha_2}{\alpha_1}
\]

A second problem is that \( E[\hat{\alpha}_2 \hat{\alpha}_1] \) may not even exist. To address both of these problems, the estimator \( \frac{\hat{\alpha}_2}{\hat{\alpha}_1} \), which will be denoted as \( f(\hat{\Theta}) \) can be approximated using a (second-order) multivariate Taylor series expansion around the estimand \( f(\Theta) \):

\[
f(\hat{\Theta}) \approx f(\Theta) + \sum_{\theta \in \Theta} (\hat{\theta} - \theta) f_\hat{\theta}(\Theta) + \frac{1}{2} \sum_{\theta \in \Theta} \sum_{\theta' \in \Theta} (\hat{\theta} - \theta)(\hat{\theta'} - \theta') f_{\hat{\theta}\hat{\theta}'}(\Theta)
\]

where \( \Theta \) contains the full set of parameters (denoted individually by \( \theta \)), \( f_\hat{\theta} \) refers to the first derivative of \( f \) with respect to \( \hat{\theta} \), and \( f_{\hat{\theta}\hat{\theta}'} \) refers to the second derivative of \( f \) with respect to \( \hat{\theta} \) and \( \hat{\theta}' \).

If we treat the higher-order terms in the Taylor series expansion as negligible, as conventionally done, then we can identify the approximate divergence between the estimator and the estimand, which is a quantity for which we can characterize the moments:

\[
E \left[ \sum_{\theta \in \Theta} (\hat{\theta} - \theta) f_\hat{\theta}(\Theta) + \frac{1}{2} \sum_{\theta \in \Theta} \sum_{\theta' \in \Theta} (\hat{\theta} - \theta)(\hat{\theta'} - \theta') f_{\hat{\theta}\hat{\theta}'}(\Theta) \right]
\]

The first-order terms in this expression are zero in expectation (i.e. \( E[\hat{\theta} - \theta] = 0 \)), while the leading components of the second-order terms are covariances in expectation (i.e. \( E[(\hat{\theta} - \theta)(\hat{\theta'} - \theta')] = \text{Cov}(\hat{\theta}, \hat{\theta'}) \)). Thus, the divergence is approximately:

\[
\frac{1}{2} \sum_{\theta \in \Theta} \sum_{\theta' \in \Theta} \text{Cov}(\hat{\theta}, \hat{\theta'}) f_{\hat{\theta}\hat{\theta}'}(\Theta)
\]

This divergence can thus be estimated—by plugging in \( \hat{\Theta} \) for \( \Theta \) and estimating the covariances—and then subtracted from the simple estimator \( f(\hat{\Theta}) \) of interest to yield an adjusted estimator that is approximately centered on the estimand of interest. Also evident from the expression
is that this divergence term goes to zero as the sample size $n$ grows to infinity. The following applies this process to the actual estimators in question.

A. Adjusted Estimators under the No-Interaction Assumption

A.1. Adjusted Estimator 1

The estimator for the first estimand is $\hat{\alpha}_2^2 \hat{\beta}_1^2 = \hat{\alpha}_2 \hat{\alpha}_1$. In expectation, the second-order Taylor Series expansion of the estimator, $T(\hat{\alpha}_2^2 \hat{\alpha}_1^2)$, around the estimand is:

$$E \left[ T \left( \frac{\hat{\alpha}_2}{\hat{\alpha}_1} \right) \right] \approx \frac{\alpha_2}{\alpha_1} - \frac{\text{Cov}(\hat{\alpha}_1, \hat{\alpha}_2)}{\alpha_1^2} + \frac{\text{Var}(\hat{\alpha}_1) \alpha_2}{\alpha_1^3}$$

Hence, we can identify the component of the approximation that diverges from the estimand. Because of the exogeneity of $T$, $\alpha_1$ and $\alpha_2$ can both be estimated without bias, allowing for the individual pieces of that component to be estimated by regression. This can then be subtracted from the estimator $\hat{\alpha}_2 \hat{\alpha}_1$ to yield an adjusted estimator approximately centered on the estimand:

$$\hat{\alpha}_2 + \hat{\text{Cov}(\hat{\alpha}_1, \hat{\alpha}_2)} \hat{\alpha}_1^2 - \hat{\text{Var}(\hat{\alpha}_1) \alpha_2} \hat{\alpha}_1$$

In the special case of balanced control and treatment assignment (i.e. $P(C) = P(T_1) = P(T_2) = \frac{1}{3}$), the adjusted estimator simplifies to:

$$\frac{\hat{\alpha}_2}{\hat{\alpha}_1} + \frac{3\hat{\sigma}_\eta^2}{\hat{\alpha}_1^2 N} - \frac{6\hat{\sigma}_\eta^2 \hat{\alpha}_2}{\hat{\alpha}_1^3 N}$$

where $\hat{\sigma}_\eta^2$ refers to the estimated error variance from equation (5). Clearly, as $N$ grows to infinity, this converges on the simple estimator $\frac{\hat{\alpha}_2}{\hat{\alpha}_1}$.

A.2. Adjusted Estimator 2

The simple estimator for the second estimand is $(\hat{\alpha}_2^2 \hat{\beta}_2) / (\hat{\alpha}_1^2 \hat{\tau}_1) = (\hat{\alpha}_2) / (\hat{\tau}_1) = \hat{\alpha}_2 \hat{\tau}_1$. As above, a second-order Taylor Series expansion can be used to formulate an adjusted estimator that is approximately centered on the estimand in finite samples:

$$\frac{\hat{\alpha}_2 \hat{\tau}_1}{\hat{\alpha}_1 \hat{\tau}_2} - \hat{\text{Var}(\hat{\alpha}_1) \hat{\alpha}_2 \hat{\tau}_1} - \hat{\text{Var}(\hat{\tau}_2) \hat{\alpha}_2 \hat{\tau}_1} + \hat{\text{Cov}(\hat{\alpha}_2, \hat{\alpha}_1) \hat{\tau}_1} \hat{\alpha}_1 \hat{\tau}_2 + \hat{\text{Cov}(\hat{\alpha}_2, \hat{\tau}_2) \hat{\tau}_1} \hat{\alpha}_1 \hat{\tau}_2$$

$$- \hat{\text{Cov}(\hat{\alpha}_2, \hat{\tau}_1)} \hat{\alpha}_1 \hat{\tau}_2 - \hat{\text{Cov}(\hat{\alpha}_1, \hat{\tau}_2) \hat{\alpha}_2 \hat{\tau}_2} + \hat{\text{Cov}(\hat{\alpha}_1, \hat{\tau}_1) \hat{\alpha}_2 \hat{\tau}_2} + \hat{\text{Cov}(\hat{\tau}_2, \hat{\tau}_1) \hat{\alpha}_2 \hat{\tau}_2}$$

vi
B. Adjusted Estimators when Relaxing the No-Interaction Assumption

Having discarded the no-interaction assumption, the estimator of the first estimand of interest, $\hat{\alpha}_{2(1)}$, is $\hat{\alpha}_{2(1)} = \hat{\alpha}_{2(1)}^{\alpha_{1}}\hat{\alpha}_{1(1)}$.

As shown, $\lim_{N \to \infty} \hat{\omega}_{j} = \omega_{j} + \xi_{j}$, because of a confounding bias that does not disappear asymptotically. Let $\omega_{j}^\ast$ denote the biased and inconsistent version of $\omega_{j}$ (i.e. $\lim_{N \to \infty} \hat{\omega}_{j} = \omega_{j}^\ast$).

As shown above, under certain reasonable and testable assumptions, $\hat{\alpha}_{2(1)}^{\alpha_{1}}\hat{\alpha}_{1(1)}$ is conservative (i.e. attenuated toward 1) for $\hat{\alpha}_{2(1)}^{\alpha_{1}}\hat{\alpha}_{1(1)}$ and hence the estimator of interest is asymptotically conservative for the estimand of interest. Unfortunately, for two reasons, this does not mean that in small samples the estimator of interest is in expectation also conservative. First, as before, the expectation may not even actually exist. Second, also as before, the ratio form of the estimand leads the estimator to be decentered from the point to which it converges. However, also as in the case with the no-interaction assumption, a second-order Taylor Series expansion can be used to construct an adjusted estimator that in finite samples is approximately centered upon the conservative point for which the estimator is consistent.

Specifically, the adjusted estimator is:

$$\frac{\hat{\alpha}_{2}\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}} - \hat{\text{Var}}(\hat{\alpha}_{1})\frac{\hat{\alpha}_{2}\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} - \hat{\text{Var}}(\hat{\omega}_{1})\frac{\hat{\alpha}_{2}\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} + \hat{\text{Cov}}(\hat{\alpha}_{2},\hat{\alpha}_{1})\frac{\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} + \hat{\text{Cov}}(\hat{\alpha}_{2},\hat{\omega}_{1})\frac{\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} - \hat{\text{Cov}}(\hat{\alpha}_{2},\hat{\omega}_{2})\frac{1}{\hat{\alpha}_{1}\hat{\omega}_{1}} - \hat{\text{Cov}}(\hat{\alpha}_{1},\hat{\omega}_{1})\frac{\hat{\alpha}_{2}\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} + \hat{\text{Cov}}(\hat{\alpha}_{1},\hat{\omega}_{2})\frac{\hat{\alpha}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}} + \hat{\text{Cov}}(\hat{\omega}_{1},\hat{\omega}_{2})\frac{\hat{\alpha}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}}$$

where $\hat{\omega}_{j} = \hat{\beta} + \hat{\gamma}_{j}$ from Equation 6 and covariance terms can be estimated via the bootstrap.

Following the same approach for the second CCM estimand, the adjusted version of the second estimator, $\left(\frac{\hat{\alpha}_{2}\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}}\right)$, is:

$$\frac{\hat{\alpha}_{2}\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}} - \hat{\text{Var}}(\hat{\alpha}_{1})\frac{\hat{\alpha}_{2}\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} - \hat{\text{Var}}(\hat{\omega}_{1})\frac{\hat{\alpha}_{2}\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} + \hat{\text{Cov}}(\hat{\alpha}_{2},\hat{\alpha}_{1})\frac{\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} + \hat{\text{Cov}}(\hat{\alpha}_{2},\hat{\omega}_{1})\frac{\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} - \hat{\text{Cov}}(\hat{\alpha}_{2},\hat{\omega}_{2})\frac{1}{\hat{\alpha}_{1}\hat{\omega}_{1}} - \hat{\text{Cov}}(\hat{\alpha}_{1},\hat{\omega}_{1})\frac{\hat{\alpha}_{2}\hat{\omega}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}} + \hat{\text{Cov}}(\hat{\alpha}_{1},\hat{\omega}_{2})\frac{\hat{\alpha}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}} + \hat{\text{Cov}}(\hat{\omega}_{1},\hat{\omega}_{2})\frac{\hat{\alpha}_{2}}{\hat{\alpha}_{1}\hat{\omega}_{1}^{2}}$$

In sum, if the assumption of no interaction between the treatments and the mediator is relaxed, the CCM estimators are no longer consistent, but they are asymptotically conservative...
provided additional conditions are met. Those additional conditions are both theoretically reasonable and empirically testable. Furthermore, finite-sample adjustments can be added to the estimators such that they are also conservative in smaller samples.
Appendix C: Tests and Sensitivity Analysis for the Conservatism of Estimators with Interactions

As explained in the main text, given the conditions described in Proposition 2, the bias involved in estimating \( \frac{\kappa_2(1)}{\kappa_1(1)} \) and \( \left( \frac{\kappa_1(1)}{\tau_2} \right) \) results in conservative (attenuated toward 1) estimates of these estimands. While assumption 4 (no interaction between the treatments and mediator) was relaxed, Proposition 2 introduces the following additional condition that was not present in Proposition 1: \( \omega_2 \xi_1 > \omega_1 \xi_2 \). This appendix shows how this condition can be partially assessed empirically.

Recall the semi-parametric model:

\[
M_i = \pi + \alpha_1 T_{1i} + \alpha_2 T_{2i} + \eta_i
\]

\[
Y_i = \lambda + \delta_1 T_{1i} + \delta_2 T_{2i} + \beta M_i + \gamma_1 T_{1i} M_i + \gamma_2 T_{2i} M_i + \iota_i
\]

\[
Y_i = \chi + \tau_1 T_{1i} + \tau_2 T_{2i} + \rho_i
\]

Now, consider equations 5 and 6 in the model by treatment subsets:

\[
(M_i | T_{1i} = 1, T_{2i} = 0) = \pi + \alpha_1 + \eta_i
\]

\[
(Y_i | T_{1i} = 1, T_{2i} = 0) = \lambda + \delta_1 + \omega_1 M_i + \iota_i
\]

\[
(M_i | T_{1i} = 0, T_{2i} = 1) = \pi + \alpha_2 + \eta_i
\]

\[
(Y_i | T_{1i} = 0, T_{2i} = 1) = \lambda + \delta_2 + \omega_2 M_i + \iota_i
\]

where \( \omega_1 = \beta + \gamma_1 \) and \( \omega_2 = \beta + \gamma_2 \). Given the saturation of the model presented in equations 5 and 6, estimation of the parameters via linear least squares regression would yield identical results if applied to equations 5 and 6 or the subsetted equations.

Consider estimation of \( \omega_1 \) and \( \omega_2 \) via linear least squares regression as applied to subsetted equations 9 and 11. For both cases, \( j = 1, 2 \), this is a bivariate regression, and thus:

\[
\text{plim}_{N \to \infty} \hat{\omega}_j = \frac{\text{Cov}(Y_i, M_i | T_{ij} = 1, T_{ij'} = 0)}{\text{Var}(M_i | T_{ij} = 1, T_{ij'} = 0)} = \frac{\text{Cov}(\lambda + \delta_j + \omega_j M_i + \iota_i, M_i | T_{ij} = 1, T_{ij'} = 0)}{\text{Var}(M_i | T_{ij} = 1, T_{ij'} = 0)}
\]

\[
= \omega_j \frac{\text{Cov}(M_i, M_i | T_{ij} = 1, T_{ij'} = 0) + \text{Cov}(\iota_i, M_i | T_{ij} = 1, T_{ij'} = 0)}{\text{Var}(M_i | T_{ij} = 1, T_{ij'} = 0)}
\]

\[
= \omega_j + \frac{\text{Cov}(\iota_i, \eta_i | T_{ij} = 1, T_{ij'} = 0)}{\text{Var}(\eta_i | T_{ij} = 1, T_{ij'} = 0)}
\]

That is,

\[
\text{plim}_{N \to \infty} \hat{\omega}_1 = \omega_1 + \xi_1 = \omega_1 + \frac{\text{Cov}(\iota_i, \eta_i | T_{i1} = 1, T_{i2} = 0)}{\text{Var}(\eta_i | T_{i1} = 1, T_{i2} = 0)}
\]

\[
\text{plim}_{N \to \infty} \hat{\omega}_2 = \omega_2 + \xi_2 = \omega_2 + \frac{\text{Cov}(\iota_i, \eta_i | T_{i1} = 0, T_{i2} = 1)}{\text{Var}(\eta_i | T_{i1} = 0, T_{i2} = 1)}
\]
Now, consider that: \( \omega_2 \xi_1 > \omega_1 \xi_2 \) implies that

\[
\left( \text{plim}_{N \to \infty} \hat{\omega}_2 - \xi_2 \right) \xi_1 > \left( \text{plim}_{N \to \infty} \hat{\omega}_1 - \xi_1 \right) \xi_2
\]

\[
\left( \text{plim}_{N \to \infty} \hat{\omega}_2 \right) \xi_1 > \left( \text{plim}_{N \to \infty} \hat{\omega}_1 \right) \xi_2
\]

\[
\left( \text{plim}_{N \to \infty} \hat{\omega}_2 \right) \frac{\text{Cov}(\iota_i, \eta_i|T_{i1} = 1, T_{i2} = 0)}{\text{Var}(\eta_i|T_{i1} = 1, T_{i2} = 0)} > \left( \text{plim}_{N \to \infty} \hat{\omega}_1 \right) \frac{\text{Cov}(\iota_i, \eta_i|T_{i1} = 0, T_{i2} = 1)}{\text{Var}(\eta_i|T_{i1} = 0, T_{i2} = 1)}
\]

Unfortunately, the possibility of unobserved confounding given non-randomization of the mediator makes it impossible to reliably estimate or compare \( \text{Cov}(\iota_i, \eta_i|T_{ij} = 1, T_{ij'} = 0) \) for \( i = 1, 2 \) without additional assumptions. However, in large samples, \( \text{plim}_{N \to \infty} \hat{\omega}_j \) can be approximated by \( \hat{\omega}_j \) and \( \text{Var}(\eta_i|T_{ij} = 1, T_{ij'} = 0) \) can be approximated by \( \text{Var}(\eta_i|T_{ij} = 1, T_{ij'} = 0) = \hat{\sigma}_{\eta_j}^2 \) using the observed data.

Hence,

\[
\left( \text{plim}_{N \to \infty} \hat{\omega}_2 \right) \frac{\text{Cov}(\iota_i, \eta_i|T_{i1} = 1, T_{i2} = 0)}{\text{Var}(\eta_i|T_{i1} = 1, T_{i2} = 0)} > \left( \text{plim}_{N \to \infty} \hat{\omega}_1 \right) \frac{\text{Cov}(\iota_i, \eta_i|T_{i1} = 0, T_{i2} = 1)}{\text{Var}(\eta_i|T_{i1} = 0, T_{i2} = 1)}
\]

can be partially assessed via:

\[
\hat{\omega}_2 \hat{\sigma}_{\eta_2}^2 > \hat{\omega}_1 \hat{\sigma}_{\eta_1}^2
\]
Appendix D: Simulations when Relaxing the No-Interaction Assumption

To illustrate the properties of the CCM estimators once the no-interaction assumption has been relaxed, this section presents the results of a simulation. The data-generating process was similar to that of the simulation presented earlier except, in this case, the effect of the mediator on the outcome involves interactions with both treatments. In addition, the simulated sample size has been increased to 1000 units per treatment condition in order to better illustrate the asymptotic tendencies\(^{15}\). As before, positive bias is introduced by construction through the omission in the estimation of a confounder that affects both the outcome and mediator. Also as before, the ACME for the treated for the second treatment is larger than that of the first treatment; further, the interaction between the mediator and the second treatment is also made larger than the interaction between the mediator and the first treatment. Thus, the additional conditions required for conservative estimation of the CCM estimands are met. Figure D1 shows the resulting estimates in the simulation.

Figure D1: Comparative Causal Mediation Simulation: With Interactions

As can be seen in the top row of Figure D1, the estimators of the ACMEs for the treated are again biased upward and, as a result, also have bad confidence-interval coverage. In contrast, however, the estimator of the ratio of ACMEs for the treated is much more well-behaved. While no longer consistent, and hence not properly centered in this medium-sized sample, the estimator is conservative (attenuated toward unity), as indicated by the mean

\(^{15}\)For this reason, the finite-sample adjustments make little difference, and hence the adjusted estimators are not presented here.
estimate being closer to one than the true value. As a result of this conservatism, there is unfortunately confidence-interval under-coverage. However, what makes this problem less concerning is that the under-coverage is the result of attenuated estimates, as shown by the majority of bad confidence intervals being below the true value, rather than the result of systematically undersized confidence intervals.

The results are similar for the bottom row of Figure D1, which presents the estimates for the proportions mediated, as well as the ratio of the proportions mediated. Again, the traditional estimators are biased upward, while the CCM estimator is conservative.
Appendix E: Application Text

Prologue

Please consider the following hypothetical scenario:
ISIS militants in Iraq were threatening rocket attacks on neighboring countries in the region. In response, the U.S. government considered taking military action. The U.S. ruled out drone strikes and other options because the ISIS militants were hiding in a civilian zone, and the U.S. government wanted to avoid harming civilians. Instead, U.S. commandos were deployed in a covert operation. In order to avoid inflicting permanent harm on nearby civilians, the commandos used a non-lethal “incapacitating” chemical gas to knock out and capture the ISIS militants. However, critics of the operation have pointed out that people have varying levels of sensitivity to the incapacitating gas, and exposure can be fatal for some people. Hence, the operation may have put civilian lives in harm’s way.

Treatment

CONTROL (no additional information provided)

OR

INFORMAL TREATMENT: Furthermore, the U.S. government has pledged never to use incapacitating chemical gas in previous public statements. Hence, the U.S. government has broken its pledge.

OR

LEGAL TREATMENT: Furthermore, the U.S. government has pledged never to use incapacitating chemical gas under its membership in the Chemical Weapons Convention, the international treaty banning chemical weapons. Hence, the U.S. government has broken international law.

DV 1: Disapproval

In general, do you approve or disapprove of the U.S. government’s decision to use the incapacitating gas in the operation?

- Approve Strongly, Approve, Neither Approve nor Disapprove, Disapprove, Disapprove Strongly

- Variable is dichotomized for analysis, with 1 indicating “Disapprove” or “Disapprove Strongly,” and 0 otherwise.

DV 2: Punishment

Imagine that one of your U.S. Senators voted in favor of using the incapacitating gas. Would this increase or decrease your willingness to vote for that Senator in the next election?

- Increase Greatly, Increase, Neither Increase nor Decrease, Decrease, Decrease Greatly

- Variable is dichotomized for analysis, with 1 indicating “Decrease” or “Decrease Greatly,” and 0 otherwise.
Mediator: Perceived Immorality

To what extent do you believe that the decision to use the incapacitating gas in the operation was morally right or wrong?

- Definitely Right, Probably Right, Not Morally Right or Wrong, Probably Wrong, Definitely Wrong

- Variable is dichotomized for analysis, with 1 indicating “Probably Wrong” or “Definitely Wrong” and 0 otherwise.
## Appendix F: Application Sample Demographics

Table F1: Sample Demographics

|                | Female  | Male    |
|----------------|---------|---------|
| Gender         | 46.3%   | 53.7%   |
|                |         |         |
| Age            |         |         |
| 18-29          | 38.9%   |         |
| 30-44          | 41.1%   |         |
| 45-64          | 18.5%   |         |
| 65+            | 1.6%    |         |
|                |         |         |
| Education      |         |         |
| No High School | 0.8%    |         |
| High School    | 11.9%   |         |
| Some College   | 34.1%   |         |
| College Graduate | 53.1% |         |