PA

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

Kirk Bansak

Assistant Professor of Political Science, University of California, San Diego, Department of Political Science, La Jolla, CA 92093, USA. Email: kbansak@ucsd.edu

Abstract

Experiments often include multiple treatments, with the primary goal to compare the causal effects of those treatments. This study focuses on 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. Further, it 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, which is a strong and nonrefutable assumption that must typically be made for consistent estimation of individual causal mediation effects. To illustrate the method, the study presents an original application investigating whether and how the international legal status of a foreign policy commitment can increase the domestic political "audience costs" that democratic governments suffer for violating such a commitment. The results provide novel evidence that international legalization can enhance audience costs via multiple causal channels, including by amplifying the perceived immorality of violating the commitment.

Keywords: causal inference, causal mediation, experimental design, experimental studies, audience costs

1 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, is often implemented in experimental research. In the most commonly used "single-experiment design," the treatment variable is randomized and the mediator(s) observed.

Another common practice in experimental research is the design of experiments featuring multiple treatment arms. As knowledge and empirical results have accumulated in various academic subfields 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. Across scientific, social scientific, and policy/program evaluation contexts, 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.

Author's note: For helpful advice, the author thanks Avidit Acharya, Justin Grimmer, Jens Hainmueller, Andy Hall, Kosuke Imai, Hye-Sung Kim, Ken Scheve, Mike Tomz, Teppei Yamamoto, and three anonymous reviewers. Replication materials are available in Bansak (2019). The author declares that he has no competing interests.

Political Analysis (2020)
vol. 28:222–243
DOI: 10.1017/pan.2019.31

Published
2 August 2019

Corresponding author
Kirk Bansak

Edited by
Jeff Gill

This study focuses on 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. Specifically, these estimands take the form of ratios between mediation effects. In addition, the value of this approach 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 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.

1.1 Related Literature

Estimation of causal mediation effects has traditionally been implemented using the parametric structural equation modeling (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, Keele, and Tingley 2010a; Imai, Jo, and Stuart 2011a; Imai *et al.* 2011b) and both parametric and nonparametric SEM frameworks (Pearl 2001; VanderWeele 2009). The parametric SEM framework has been critiqued in particular for its inflexibility and reliance on functional form assumptions, with researchers instead advocating for more generalized, nonparametric formulations of causal mediation effects (Imai, Keele, and Tingley 2010a; Imai *et al.* 2011b; Pearl 2001, 2014).[1]

This study employs the potential outcomes formalization of causal mediation effects presented by Imai, Keele, and Tingley (2010a), Imai, Keele, Yamamoto (2010b). In addition, to formulate the methods, this study adapts the semiparametric model introduced by Imai and Yamamoto (2013), which presents a convenient and interpretable statistical structure yet also avoids the rigidity of the traditional parametric SEM framework by allowing for unit-specific parameters. In addition, this flexibility allows for the causal mediation effects as defined using potential outcomes notation to be easily expressed within the model. For other semiparametric modeling approaches to causal mediation analysis, see Glynn (2012) and Tchetgen and Shpitser (2012).

This study also follows much of the methodological literature on causal mediation preceding it in terms of key assumptions that are employed. A version of the assumption of no interaction between treatment and mediator, which was introduced and formalized to identify mediation effects in earlier work on causal mediation (Robins and Greenland 1992; Robins 2003), is employed for some of the results in this study. However, as emphasized by Robins (2003) and Imai, Tingley, and Yamamoto (2013), the no-interaction assumption must generally hold at the individual level for existing causal mediation methods, whereas this assumption must simply hold on average in the comparative context introduced in this study. Following previous work (Imai, Keele, Yamamoto 2010b; Kraemer *et al.* 2008; Imai and Yamamoto 2013), this study also presents results when the no-interaction assumption is relaxed. In addition, the assumption of no covariance between (individual-level) causal parameters is employed in this study. As has been highlighted by Hong (2015, chapter 10), this assumption is routinely employed (or implied by other assumptions) in existing approaches to causal mediation analysis.

1 See Shpitser and VanderWeele (2011) and VanderWeele (2015) for a discussion of the connection between the nonparametric SEM and potential outcomes approaches to causal mediation analysis.

While continuing to utilize certain assumptions, a key contribution of this study is in allowing for a relaxation of the assumption of no unobserved confounding of the mediator–outcome relationship. Loeys *et al.* (2016) make a similar contribution of highlighting how certain causal mediation quantities of interest can still be identified when relaxing this assumption. Specifically, Loeys *et al.* (2016) show how an "index for moderated mediation," which measures the extent to which a causal mediation effect varies by the level of other variables (moderators), can be identified under certain conditions without the assumption of no unobserved mediator–outcome confounding. In contrast to the present study, however, the structural framework used by Loeys *et al.* (2016) employs constant effects rather than unit-specific parameters.

It 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 each mediator to a different degree 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. 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).

In addition, another related line of research has focused on identification and estimation of "controlled direct effects," which refer to the direct effect of a treatment when fixing the mediator at a common value for all units, in contrast to "natural direct effects," which fix the mediator at unit-specific potential values for each unit under a particular treatment level, such as under nonexposure (e.g. Robins 1997; Pearl 2001; VanderWeele 2014; Acharya, Blackwell, and Sen 2016). Controlled and natural direct effects are not considered in this study. Guidance on the difference between these two types of direct effects, their relationship with causal mediation effects, and how to identify and estimate average controlled direct effects can be found in Acharya, Blackwell, and Sen (2016).

1.2 Outline

The remainder of this study is organized as follows. Section 2 provides motivation and explains the value, in both theoretical and policy contexts, for comparing the causal mediation effects of multiple treatments. Section 3 formally introduces the new CCM estimands. Section 4 then presents an estimation strategy, describing the assumptions and methods under which the CCM estimands can be estimated consistently. Section 5 presents simulations to illustrate the properties of the estimators. Section 6 then describes how these properties change—namely, how the CCM estimands can be estimated conservatively but no longer consistently—under a relaxed set of assumptions. To illustrate the CCM method, Section 7 presents an original application, investigating the effect of international legality on the domestic political costs that democratic governments suffer for violating foreign policy commitments. Section 8 concludes.

2 Motivation for Comparing Causal Mediation Effects

In experimental research contexts involving multiple related treatments, theories on why one treatment should have a larger effect than another are linked to the presumed mechanism(s) through which each treatment propagates its effect. As a prelude to the application presented later in this study, consider the recent accumulation of experimental evidence in the political science literature on "audience costs" (for a brief review, see Hyde 2015).[2] These many studies have

2 Audience costs refer to the electoral costs to politicians (i.e. punishment by voters) for breaking policy commitments. The past decade has 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).

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 audience costs. For instance, a legalized foreign policy commitment could gain audience cost strength over an informal commitment via various mechanisms, such as a heightened sense of immorality for violating legalized commitments on the part of citizens, or a belief that violating legalized commitments is more likely to lead to international retaliation.

Another example exists in the literature on party cues in American politics, which includes a wealth of experimental studies that investigate party cue effects on voter attitudes and behavior (e.g. Kam 2005; Arceneaux 2008).[3] As these studies have highlighted, there are various types of party cues, and 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, Federico, and Kittilson 2009; Nicholson 2012). There may be various mechanisms by which out-party cue effects can exceed those of in-party cues—for instance, the possibility that out-party cues elicit stronger emotional reactions than in-party cues, or the possibility that out-party cues may actually be more informative than in-party cues. Such possibilities could be tested by rigorously comparing the mechanisms underlying each set of party cues.

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. 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. Ideally, however, the effectiveness of any preferred policy intervention should be generalizable across time and different localities. One important means of assessing generalizability is to develop a comprehensive understanding of the mechanisms underlying different treatments.

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 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, and a superficial interpretation of these results would then be that the two programs are interchangeable. However, to enable more efficient policy targeting, it would be useful to investigate the causal mechanism differences between the two job training programs, as 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, depending upon local economic conditions.

3 Comparative Causal Mediation (CCM) Estimands

As a frame of reference, consider the single-treatment experimental setting. 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). Following the potential outcomes approach to causal mediation analysis presented by Imai, Keele, and Tingley (2010a), Imai, Keele, Yamamoto (2010b),

3 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.

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. The ACME 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 related 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 $M(t_1, t_2)$. 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 $t_2 = 1$. This allows for defining a separate $ACME_j$ and ATE_j for each treatment T_j as follows:

$$ACME_1 = \kappa_1(t_1) = E[Y(t_1, 0, M(1, 0)) - Y(t_1, 0, M(0, 0))] \quad (1)$$

$$ATE_1 = \tau_1 = E[Y(1, 0, M(1, 0)) - Y(0, 0, M(0, 0))] \quad (2)$$

$$ACME_2 = \kappa_2(t_2) = E[Y(0, t_2, M(0, 1)) - Y(0, t_2, M(0, 0))] \quad (3)$$

$$ATE_2 = \tau_2 = E[Y(0, 1, M(0, 1)) - Y(0, 0, M(0, 0))]. \quad (4)$$

Note that all effects ($ACME$s and ATEs) are referenced against the pure control condition.

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 CCM estimands of interest.

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

$$Estimand\ 1: \frac{ACME_2}{ACME_1} = \frac{\kappa_2(t_2)}{\kappa_1(t_1)} \quad Estimand\ 2: \frac{\left(\frac{ACME_2}{ATE_2}\right)}{\left(\frac{ACME_1}{ATE_1}\right)} = \frac{\left(\frac{\kappa_2(t_2)}{\tau_2}\right)}{\left(\frac{\kappa_1(t_1)}{\tau_1}\right)}.$$

The first estimand measures the extent to which one treatment has a stronger causal mediation effect transmitted via the mediator of interest relative to the other treatment. In contrast, the second estimand measures the extent to which one treatment has a greater proportion of its total effect transmitted through the mediator of interest relative to the other treatment, which allows for testing the extent to which the mediator is more important to the overall causal anatomy of one treatment. For additional discussion on the types of research questions and hypotheses each estimand is better suited to address, see Appendix H.

4 Estimation of Comparative Causal Mediation

4.1 Model

Consider a simple random sample of N observations. Let $Y_i(t_1, t_2, m)$ and $M_i(t_1, t_2)$ denote the potential outcomes for unit i. Let T_{1i} (T_{2i}) denote the first (second) treatment indicator, which equals one if unit i receives the first (second) treatment and zero otherwise. The observed

mediator M_i equals $M_i(T_{1i}, T_{2i})$, and the observed outcome Y_i equals $Y_i(T_{1i}, T_{2i}, M_i(T_{1i}, T_{2i}))$. Note that given the mutual exclusivity of the two binary treatments, $Y_i(1, 1, m)$ and $M_i(1, 1)$ do not exist.

Adapting the semiparametric model introduced by Imai and Yamamoto (2013), the potential outcomes are modeled using the following structural equations:

$$M_i(t_1, t_2) = \pi_i + \alpha_{1i} t_1 + \alpha_{2i} t_2$$
$$Y_i(t_1, t_2, m) = (\lambda_i + \delta_{1i} t_1 + \delta_{2i} t_2) + (\beta_i + \gamma_{1i} t_1 + \gamma_{2i} t_2) m.$$

The model shares some basic notational similarities with the parametric structural equation models often used to describe causal mediation, though a key difference is that the equations here allow for unit-specific parameters. The relationships implicitly assume that the potential outcomes are linear in m, but are otherwise flexible given mutually exclusive, binary treatments and the unit-specific parameters. In the case of a binary mediator, the relationships become fully flexible and nonparametric. This semiparametric setup highlights the relationship between the ACME as defined under the potential outcomes approach and the natural indirect effect as defined by structural equation models of causal mediation:

$$ACME_1 = \kappa_1(t_1) = E[Y_i(t_1, 0, M_i(1, 0)) - Y_i(t_1, 0, M_i(0, 0))] = E[\alpha_{1i}(\beta_i + \gamma_{1i} t_1)]$$
$$ACME_2 = \kappa_2(t_2) = E[Y_i(0, t_2, M_i(0, 1)) - Y_i(0, t_2, M_i(0, 0))] = E[\alpha_{2i}(\beta_i + \gamma_{2i} t_2)].$$

In the classic SEM framework (Baron and Kenny 1986), constant effects and no interaction between treatment and mediator are assumed. Applying those assumptions to the two-treatment context here yields $E[\alpha_{ji}(\beta_i + \gamma_{ji} t_j)] = \alpha_j \beta$, where $j = 1, 2$ denotes the treatment, which is indeed the classic product-of-coefficients result in the SEM framework.[4] However, this study will not assume constant effects, and a no-interaction assumption will be introduced but then relaxed later.

In addition, define the reduced-form version of the potential outcome $Y_i(t_1, t_2, M_i(t_1, t_2)) = Y_i(t_1, t_2) = \chi_i + \tau_{1i} t_1 + \tau_{2i} t_2$, which is fully flexible given mutually exclusive, binary treatments.[5] The average treatment effects (ATEs) can thus be expressed:[6]

$$ATE_1 = \tau_1 = E[Y_i(1, 0) - Y_i(0, 0)] = E[\tau_{1i}]$$
$$ATE_2 = \tau_2 = E[Y_i(0, 1) - Y_i(0, 0)] = E[\tau_{2i}].$$

Now, following Imai and Yamamoto (2013), the unit-specific parameters can be decomposed into their means and deviations. That is, for each parameter θ_i, define $\theta = E[\theta_i]$ and $\tilde{\theta}_i = \theta_i - \theta$. This yields the following set of estimating equations where the individual-level heterogeneity is subsumed into the error terms:

$$M_i = \pi + \alpha_1 T_{1i} + \alpha_2 T_{2i} + \eta_i \tag{5}$$
$$Y_i = \lambda + \delta_1 T_{1i} + \delta_2 T_{2i} + \beta M_i + \gamma_1 T_{1i} M_i + \gamma_2 T_{2i} M_i + \iota_i \tag{6}$$
$$Y_i = \chi + \tau_1 T_{1i} + \tau_2 T_{2i} + \rho_i \tag{7}$$

4 The 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 (Jo 2008; Pearl 2014). This includes the potential outcomes framework, where it has previously been shown that the ACME is equivalent to $\alpha\beta$ under certain conditions (Imai, Keele, Yamamoto 2010b).

5 This reduced-form presentation is also employed in the single-treatment context by Glynn (2012).

6 As shown in the single-treatment context (e.g. Imai, Keele, Yamamoto 2010b), the ATEs can also be equivalently defined with reference to the full potential outcomes $Y_i(t_1, t_2, m)$ and $M_i(t_1, t_2)$ as such:

$$ATE_1 = E[Y_i(1, 0, M_i(1, 0)) - Y_i(0, 0, M_i(0, 0))]$$
$$ATE_2 = E[Y_i(0, 1, M_i(0, 1)) - Y_i(0, 0, M_i(0, 0))].$$

where $\eta_i = \tilde{\pi}_i + \tilde{\alpha}_{1i}T_{1i} + \tilde{\alpha}_{2i}T_{2i}$, $\iota_i = \tilde{\lambda}_i + \tilde{\delta}_{1i}T_{1i} + \tilde{\delta}_{2i}T_{2i} + \tilde{\beta}_i M_i + \tilde{\gamma}_{1i}T_{1i}M_i + \tilde{\gamma}_{2i}T_{2i}M_i$, and $\rho_i = \tilde{\chi}_i + \tilde{\tau}_{1i}T_{1i} + \tilde{\tau}_{2i}T_{2i}$.

4.2 Assumptions

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

To be explicit, the linearity assumption is also reiterated.

ASSUMPTION 2. Linear relationships between the potential outcomes and the mediator.

$$Y_i(t_1, t_2, m) = (\lambda_i + \delta_{1i}t_1 + \delta_{2i}t_2) + (\beta_i + \gamma_{1i}t_1 + \gamma_{2i}t_2)m.$$

As already described above, while the assumption of linearity seems demanding, it is made trivial by the employment of a binary mediator. Given a binary mediator and the two mutually exclusive binary treatments, the potential outcome model described above is fully saturated and hence "inherently linear" (Angrist and Pischke 2009, 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 model.

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.

Let N_1 denote the number of units assigned to treatment 1, N_2 the number assigned to treatment 2, and $N - N_1 - N_2$ the number assigned to the control condition (neither treatment 1 nor treatment 2). Then for any unit i,

$$P(T_{1i} = 1, T_{2i} = 0) = \frac{N_1}{N} \quad P(T_{1i} = 0, T_{2i} = 1) = \frac{N_2}{N}$$
$$P(T_{1i} = 0, T_{2i} = 0) = \frac{N - N_1 - N_2}{N} \quad P(T_{1i} = 1, T_{2i} = 1) = 0.$$

The third assumption is no treatment–mediator interactions 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 $Y_i = \lambda + \delta_1 T_{1i} + \delta_2 T_{2i} + \beta M_i + \iota_i$. The no-interaction assumption was introduced and formalized to identify the ACME in earlier literature on causal mediation analysis (Robins and Greenland 1992; Robins 2003), and it has since been commonly employed to identify the ACME in the single-treatment context. However, as emphasized by Robins (2003) and Imai, Tingley, and Yamamoto (2013), the no-interaction assumption must generally hold at the individual level in the standard single-treatment context. In contrast, here the assumption must simply hold on average. Nonetheless, compared to

assumptions 2 and 3, the no-interaction assumption is more stringent and cannot be guaranteed by design. For this reason, this assumption will be relaxed later (γ_1 and γ_2 will be allowed to be nonzero), and diagnostics will be presented to allow for an empirical assessment of the assumption.

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

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

$$Cov(\alpha_{1i}, \beta_i) = Cov(\alpha_{1i}, \gamma_{1i}) = 0$$
$$Cov(\alpha_{2i}, \beta_i) = Cov(\alpha_{2i}, \gamma_{2i}) = 0.$$

This type of no-covariance assumption is also made, implicitly or explicitly, in other approaches to causal mediation (Hong 2015). For instance, 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.[7] See Hong (2015, chapter 10) for a comprehensive overview of the no-covariance assumption as used in the various statistical approaches to causal mediation analysis. 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 α_{ji} and β_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.

4.3 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, Green, and Ha 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, Keele, Yamamoto 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 β via linear least squares regression results in the bias term $E[\hat{\beta} - \beta] = \frac{cov(\eta_i, \iota_i)}{var(\eta_i)}$. In contrast, α_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 ACME estimates.

7 As Imai, Keele, Yamamoto (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, α_i is a function of the potential values of the mediator, while β_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 α_i and β_i.

By corollary, this means 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.

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 from a larger population. Given assumptions 1–5, then the following estimators converge in probability to the estimands of interest under the usual generalized linear regression regularity conditions:*[8]

$$\operatorname*{plim}_{N\to\infty}\left(\frac{\hat{\alpha}_2^N\hat{\beta}^N}{\hat{\alpha}_1^N\hat{\beta}^N}\right)=\frac{\kappa_2(t_2)}{\kappa_1(t_1)} \quad \textit{and} \quad \operatorname*{plim}_{N\to\infty}\left(\frac{\left(\frac{\hat{\alpha}_2^N\hat{\beta}^N}{\hat{\tau}_2^N}\right)}{\left(\frac{\hat{\alpha}_1^N\hat{\beta}^N}{\hat{\tau}_1^N}\right)}\right)=\frac{\left(\frac{\kappa_2(t_2)}{\tau_2}\right)}{\left(\frac{\kappa_1(t_1)}{\tau_1}\right)}.$$

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

4.4 Scope Conditions and Issues in Ratio Estimation

A number of issues have long been noted with the use and interpretation of ratio estimators,[9] and the estimators proposed here are no exception. In particular, their ratio form has important implications for the scope conditions under which they are useful and reliable, their small-sample tendencies, uncertainty estimation, and statistical power. These issues are discussed below.

4.4.1 Scope Conditions

In addition to the obvious precondition of an experimental design featuring multiple treatments, there are other key scope conditions that dictate when the CCM methods will be usable or useful. First, each estimand is only useful when both the numerator and denominator can be estimated as having the same sign and with sufficient statistical precision. This is, first and foremost, a conceptual precondition as the estimands are conceptually meaningful and interpretable only when the ACMEs for both treatments are presumed to be nonzero in the same direction. In addition, this is also an important statistical consideration. Indeed, it has long been known that ratio estimators exhibit finite-sample distributional behavior that is difficult to formally characterize (except under special conditions) and has important implications for their central tendencies and dispersion (e.g. Fieller 1954).

Given their ratio form, the CCM estimators presented in this study share the same fundamental problem of potentially "dividing by zero" as that of weak instruments in instrumental variables (IV) estimation (Nelson and Startz 1990). Research over the past two decades to develop best practices for detecting weak instruments is thus informative here (see Andrews, Stock, and Sun (2019) for an overview). Earlier research on the matter provided the rule-of-thumb recommendation, which continues to be widely used, that IV estimates for a single endogenous regressor be considered reliable only when tests of the first-stage regression yield an F statistic greater than 10 (Staiger and Stock 1997; Stock, Wright, and Yogo 2002), and more recent research has highlighted that this simple decision rule provides relatively reliable guidance in the single-instrument case (Stock and Yogo 2005; Olea and Pflueger 2013; Andrews, Stock, and Sun 2019). Given that single-instrument IV estimation is a simple ratio estimator itself, this rule of thumb thus provides useful scope conditions for the CCM estimators as well. To implement this decision rule, first note that the two CCM estimators can be re-expressed as $\frac{\hat{\alpha}_2^N}{\hat{\alpha}_1^N}$ and $\frac{\hat{\alpha}_2^N\hat{\tau}_1^N}{\hat{\alpha}_1^N\hat{\tau}_2^N}$. For either estimator, denote the denominator

8 Proofs of propositions can be found in Appendix A.

9 For a useful summary of early results and thinking on ratio estimators, see Flueck and Holland (1976).

by $\hat{\theta}_d$, and consider the estimator unreliable if the following statistic is less than 10:

$$F = \frac{\hat{\theta}_d^2}{\widehat{Var}(\hat{\theta}_d)}.$$

Third, the estimands are also likely to be most useful when the two treatments themselves have nonzero treatment effects of the same sign as the ACMEs, and where one treatment does not clearly dominate the other. This is again a matter of both conceptual clarity and statistical properties. Conceptually, there may be limited theoretical or practical insights to be gained from comparing the mediation effects if one treatment is orders of magnitude larger than the other. This should generally not be the case, however, in the context of comparing closely related treatments, which is the motivating context for the CCM methods. In addition, note that the treatment effect estimate $\hat{\tau}_2^N$ is a component of the denominator in the second estimator and hence covered by the decision rule presented above.

4.4.2 *Finite-Sample Adjustments*

Even in the case where the scope conditions above are met, the CCM estimators are not exactly centered on the true estimand in finite samples due to their ratio form. This divergence becomes negligible as the sample size grows, and in smaller samples, finite-sample adjustments can be made. One simple and well-established method of deriving finite-sample corrections for estimators of functions, such as ratio estimators, involves Taylor series expansions (e.g. Cochran 1963; Withers 1987; Lehmann and Casella 2006, chapter 6). In this vein, Appendix B presents adjusted estimators for both CCM estimands that include finite-sample corrections derived using Taylor series expansion. Simulations, presented below, compare the adjusted estimators over the simple estimators in small samples.

4.4.3 *Uncertainty Estimation*

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 derived in some special cases (e.g. Fieller 1954) and approximation techniques based on the bootstrap developed 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 can yield approximately correct coverage under the reasonable condition that the confidence interval is actually bounded at the desired α level, which is met when the $1 - \alpha$ confidence interval of the denominator does not contain zero (Franz 2007).[10] This should be met by the scope conditions presented above, which will provide for estimator denominators that are sufficiently bounded away from zero and hence allow for the use of standard methods of confidence-interval construction, such as the Delta Method and bootstrap techniques.

10 As in general, a sufficiently large sample size is also necessary for analytic methods that rely on the central limit theorem, and for bootstrap methods to adequately approximate the population distribution.

4.4.4 *Power*

As observed by researchers of causal mediation analysis, there is a relative dearth of general methods to compute power and sample size requirements for causal mediation estimators (Fairchild and McDaniel 2017; VanderWeele 2015, chapter 7). One exception is a study by Fritz and MacKinnon (2007), which provides a table of basic power and sample size requirements based on simulations. However, given the limited number of specifications considered, these results do not allow researchers to compute power or sample size requirements for their own specific scenarios. In the CCM context, there is additional complexity in computing power given the ratio functional form and the additional parameters to estimate.

One recommended method of proceeding with a power analysis in the context of complex causal mediation models is to employ customized Monte Carlo simulations (Thoemmes, MacKinnon, and Reiser 2010; Zhang 2014; Fairchild and McDaniel 2017). In particular, Zhang (2014) presents a simulation-based method using bootstrap inference that can be adapted to the CCM estimators by simulating the model equations (5)–(7). Under the no-interaction assumption, only equations (5) and (7) would need to be simulated given how $\hat{\beta}^N$ drops out of the estimators. As generally the case in power analyses, implementation would require hypothesized parameter values and variance estimates, in this case the variance of the error terms, which could be obtained from previous or pilot studies.[11] The power to reject the null hypothesis that either estimand equals 1 at a specific level of confidence could then be computed for a given sample size, or the required sample size could be determined to achieve a desired level of power. See Zhang (2014) for systematic instructions on implementation.

5 Simulations

To illustrate the properties of the CCM method, this section presents a simulation.[12] Simulated causal mediation data were generated according to the following model, with the output of the first equation feeding into the second equation:

$$M_i = \pi_i + \alpha_{1i} T_{1i} + \alpha_{2i} T_{2i} + \psi_i X_i$$
$$Y_i = \lambda_i + \delta_{1i} T_{1i} + \delta_{2i} T_{2i} + \beta_i M_i + \phi_i X_i$$

T_1 and T_2 are indicator variables that 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 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 α_j and β 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 α_1, α_2, and β—as well as τ_1 and τ_2—were estimated by linear least squares regression according to equations (5)–(7) with γ_1 and γ_2 assumed to be zero. Thus X was omitted from the estimation process, simulating unobserved confounding.

11 The intended treatment assignment structure could then be simulated to generate values of the mediator via equation (5) and then generate outcome values using equation (7). If relaxing the no-interaction assumption, outcome values would need to be generated via equation (6).

12 Replication materials are available in Bansak (2019).

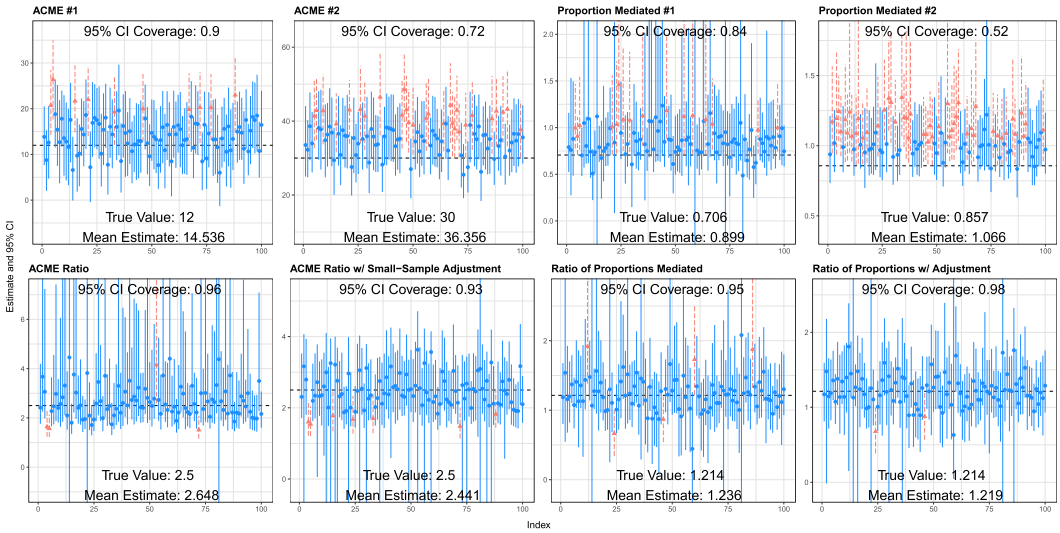

Figure 1. Comparative Causal Mediation Simulation, Without Interactions.

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 solid lines denote confidence intervals that cover the true value, whereas the dashed lines denote lack of coverage. The panels in the top row correspond to the traditional causal mediation estimands: $ACME_1$ ($E[\alpha_{1i}\beta_i]$), $ACME_2$ ($E[\alpha_{2i}\beta_i]$), proportion of ATE_1 mediated $\left(\frac{E[\alpha_{1i}\beta_i]}{E[\tau_{1i}]}\right)$, and proportion of ATE_2 mediated $\left(\frac{E[\alpha_{2i}\beta_i]}{E[\tau_{2i}]}\right)$. 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, 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 undercoverage 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 good confidence-interval coverage (subject to simulation error). In addition, the small-sample adjustments slightly improve the mean estimates, but in doing so they also substantially inflate the variance and increase the number of confidence intervals that blow up below zero from 3 to 18. The results are the same in the bottom right two panels, which show the simple and adjusted estimators for the ratio of proportions mediated. Again, the small-sample adjustments slightly improve the mean estimates at the cost of inflated variance, and an increase in the number of confidence intervals that blow up below zero from 4 to 8.

6 Relaxing the No-Interaction Assumption

6.1 Setup

Following Imai and Yamamoto (2013), the semiparametric model presented earlier, equations (5)–(7), can proceed without assumption 4 and hence allow 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:

$$\textit{Estimand } 1: \frac{\kappa_2(1)}{\kappa_1(1)} = \frac{E[\alpha_{2i}\omega_{2i}]}{E[\alpha_{1i}\omega_{1i}]} \quad \textit{Estimand } 2: \frac{\left(\frac{\kappa_2(1)}{\tau_2}\right)}{\left(\frac{\kappa_1(1)}{\tau_1}\right)} = \frac{\left(\frac{E[\alpha_{2i}\omega_{2i}]}{E[\tau_{2i}]}\right)}{\left(\frac{E[\alpha_{1i}\omega_{1i}]}{E[\tau_{1i}]}\right)}.$$

6.2 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 $\hat{\tau}_2$, $\hat{\tau}_1$, $\hat{\alpha}_2$, and $\hat{\alpha}_1$ under standard regularity conditions, but not for $\hat{\beta}$, $\hat{\gamma}_2$, and $\hat{\gamma}_1$.[13] 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 $\left(\frac{\hat{\alpha}_2(\hat{\beta}+\hat{\gamma}_2)}{\hat{\tau}_2}\right)\Big/\left(\frac{\hat{\alpha}_1(\hat{\beta}+\hat{\gamma}_1)}{\hat{\tau}_1}\right)$ are not consistent estimators of $\frac{\kappa_2(1)}{\kappa_1(1)}$ and $\left(\frac{\kappa_2(1)}{\tau_2}\right)\Big/\left(\frac{\kappa_1(1)}{\tau_1}\right)$, 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* ξ_1 *and* ξ_2 *the asymptotic bias components of* $\hat{\omega}_1^N$ *and* $\hat{\omega}_2^N$*, respectively (i.e.* $\operatorname{plim}_{N\to\infty} \hat{\omega}_1^N - \omega_1 = \xi_1$ *and* $\operatorname{plim}_{N\to\infty} \hat{\omega}_2^N - \omega_2 = \xi_2$*). Make assumptions 1, 2, 3, and 5. Then, given* $\omega_2\xi_1 > \omega_1\xi_2$*, the following holds:*

$$\operatorname*{plim}_{N\to\infty} \frac{\hat{\alpha}_2^N\hat{\omega}_2^N}{\hat{\alpha}_1^N\hat{\omega}_1^N} < \frac{\kappa_2(1)}{\kappa_1(1)}$$
$$\operatorname*{plim}_{N\to\infty} \frac{\left(\frac{\hat{\alpha}_2^N\hat{\omega}_2^N}{\hat{\tau}_2^N}\right)}{\left(\frac{\hat{\alpha}_1^N\hat{\omega}_1^N}{\hat{\tau}_1^N}\right)} < \frac{\left(\frac{\kappa_2(1)}{\tau_2}\right)}{\left(\frac{\kappa_1(1)}{\tau_1}\right)}.$$

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

13 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.

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. Note that while assumption 4 was relaxed, Proposition 2 introduces the following additional condition: $\omega_2\xi_1 > \omega_1\xi_2$. As shown in Appendix C, this condition can be partially assessed empirically.

6.3 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 when the no-interaction assumption has been relaxed.

7 Application: International Law and Audience Costs

7.1 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, Rocke, and Barsoom 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, Ropp, and Sikkink 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 nondemocracies—are generally referred to as domestic "audience costs" (Fearon 1994; Morrow 2000; Tomz 2007; Weeks 2008; Jensen 2003). The types of foreign policy commitments that have been investigated in this literature vary widely. This includes commitments targeted at a purely domestic audience, such as promises by national leaders to their constituents not to engage in certain behavior or activities. This also includes commitments directed at other countries, such as threats made against aggressor countries and promises to aid allies in the event of conflict. Finally, this also includes legally formalized international commitments, such as agreements codified in treaties.

The application presented here focuses on the legal dimension of foreign policy commitments and its relationship with audience costs. An important gap remains in the relevant 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, does the dimension of

international legality actually enhance audience costs, and if it does, to what extent and why is that 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 implication 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 2000; Simmons and Hopkins 2005), it has not been explicitly tested.

7.2 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.[14] Appendix E provides the survey instrument text and variable coding rules. Appendix F provides sample demographic distributions and balance statistics across treatment conditions. Tests of the relationship between the treatment assignment and demographic covariates fail to reject the null hypothesis of independence at the 0.05 significance level, indicating good balance.

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 nonlethal 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.[15] As a result, it was possible to effectively intervene upon respondents' knowledge of the legal status of these chemical incapacitants.

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 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

14 This research was approved by the Institutional Review Board at Stanford University (Protocol 31139).

15 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).

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 outcome variables of interest. The first measured the extent to which respondents (dis)approved of the policy decision to use chemical incapacitants, and the second measured the extent to which respondents would be likely to vote for a U.S. Senator who supported the policy decision.[16] Both variables were measured in the survey on a five-point scale. To allow for easier interpretation, the analysis presented here employs dichotomized versions of these variables: whether or not the respondent disapproved, which will be called Disapproval, and whether or not the respondent would be less likely to vote for a supportive U.S. Senator, 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, respondents' perceptions of the (im)morality of the decision to use chemical incapacitants were also measured and investigated as a mediator. Normative or moral aversion represents one possible mechanism that could lead violations of international commitments, whether legalized or not, to result in public disapproval. Previous research has highlighted and tested a variety of possible mechanisms, including morality, whereby international law may affect public opinion (Chilton 2014; Chilton and Versteeg 2016). The application presented here focuses specifically on the morality mechanism because perceptions of immorality represent one of the earliest theoretical reasons noted by international relations scholars of international law that voters would more strongly disapprove of violations of legalized foreign policy commitments versus similar nonlegalized commitments (Abbott and Snidal 2000). In addition, Appendix G presents additional analysis that probes into a second possible mechanism: concerns that other countries would follow suit in developing or using chemical incapacitants and hence harm U.S. security in the long run. Other possible mechanisms that could also be active in the international security context but were not tested include fear of more immediate international retaliation or enforcement, beliefs about the efficacy of prohibited actions or behaviors, and concerns about impact on national reputation.

To test the morality mechanism, a mediator variable was constructed by asking respondents about the degree to which they believed the policy decision to use chemical incapacitants was morally right or wrong. Similar to the dependent variables, this mediator was measured on a five-point scale, and it is dichotomized to facilitate interpretation in the analysis. The binary version of the mediator captures 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 causal mediation analysis is that, even with pretreatment 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.

16 The decision was made to focus on punishment of senators rather than the president under the assumption that this would decrease the amount of partisan priming respondents were exposed to, thereby allowing for better and less contaminated measurement of their attitudes toward the scenario.

Table 1. Sample Estimates of ATEs.

	DV: Disapproval		
	$\widehat{ATE}_1$	$\widehat{ATE}_2$	$\widehat{ATE}_2 - \widehat{ATE}_1$
	Informal treatment effect	**Legal treatment effect**	**Difference in treatment effects**
Estimate	0.195	0.320	0.125
95% CI	[0.140, 0.250]	[0.263, 0.375]	[0.065, 0.185]
	DV: Punishment		
	$\widehat{ATE}_1$	$\widehat{ATE}_2$	$\widehat{ATE}_2 - \widehat{ATE}_1$
	Informal treatment effect	**Legal treatment effect**	**Difference in treatment effects**
Estimate	0.182	0.281	0.099
95% CI	[0.128, 0.235]	[0.226, 0.336]	[0.040, 0.158]

Table 2. Comparative Causal Mediation via Perceived Immorality Mechanism.

	DV: Disapproval			
	$\widehat{ACME}_1$	$\widehat{ACME}_2$	$\frac{\widehat{ACME}_2}{\widehat{ACME}_1}$	$\left(\frac{\widehat{ACME}_2}{\widehat{ATE}_2}\right)/\left(\frac{\widehat{ACME}_1}{\widehat{ATE}_1}\right)$
	Mediation Effect for Informal Treatment	**Mediation Effect for Legal Treatment**	**Ratio of Mediation Effects**	**Ratio of Proportions Mediated**
Estimate	0.113	0.177	**1.563**	**0.952**
95% CI	[0.076, 0.151]	[0.139, 0.215]	**[1.190, 2.207]**	**[0.749, 1.211]**
	DV: Punishment			
	$\widehat{ACMET}_1$	$\widehat{ACMET}_2$	$\frac{\widehat{ACMET}_2}{\widehat{ACMET}_1}$	$\left(\frac{\widehat{ACMET}_2}{\widehat{ATE}_2}\right)/\left(\frac{\widehat{ACMET}_1}{\widehat{ATE}_1}\right)$
	Mediation Effect for Informal Treatment	**Mediation Effect for Legal Treatment**	**Ratio of Mediation Effects**	**Ratio of Proportions Mediated**
Estimate	0.096	0.176	**1.829**	**1.184**
95% CI	[0.063, 0.131]	[0.135, 0.218]	**[1.329, 2.701]**	**[0.904, 1.593]**

7.3 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 1, 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.

More importantly in the context of this study, however, the results of the CCM analysis also provide 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 2 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 $\omega_2\xi_1 > \omega_1\xi_2$.[17]

Table 2 presents the causal mediation results, including the estimates of each treatment's mediation effect transmitted via the morality mechanism as well as the CCM effects. Note that the individual $\widehat{ACME}$ estimates should not be interpreted at face value themselves as they are used specifically as inputs for the CCM estimators and are likely to be individually biased and inconsistent. In contrast, under the assumptions presented in this study, the CCM estimates (presented in bold) can be interpreted. Given the large sample size, these estimates were obtained using the simple estimators,[18] and the 95% confidence intervals were computed via the nonparametric percentile bootstrap. As can be seen, the estimates of the ratio of mediation effects, $\frac{\widehat{ACME}_2}{\widehat{ACME}_1}$, 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 estimates of the ratio of proportions mediated, $\left(\frac{\widehat{ACME}_2}{\widehat{ATE}_2}\right)\Big/\left(\frac{\widehat{ACME}_1}{\widehat{ATE}_1}\right)$, 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.

In combination, these results suggest that Perceived Immorality is 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 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.[19] Appendix G presents the results when analyzing the variables on their raw five-point scale. While on a different scale, the results remain substantively and statistically unchanged.

7.4 Discussion

In addition to illustrating the CCM methods, the results of this application also contribute to the literature on audience costs. As described above, the results add to the recent accumulation of experimental evidence that reneging on foreign policy commitments can indeed substantially decrease approval of the policy decision in question. The ATEs estimated in this application, of approximately 20 to 30 percentage points greater disapproval, are substantively large and consistent in magnitude with the higher end of effects detected in previous experimental research on audience costs.[20]

In addition, this application makes a more novel contribution in specifically distinguishing between audience costs effects when the violated commitment is legalized versus not legalized. The roughly 10- to 13-percentage-point boost attributable to legalization in this application provides new evidence on the extent to which legalization enhances audience costs. Furthermore, the CCM results provide support for the theory that international legalization enhances audience costs specifically by amplifying the perceived immorality of violating the commitment. However, the results also suggest that this is not the only mechanism by which legalization enhances

17 As explained in Appendix C, this is tested partially by verifying that $\hat{\omega}_2\widehat{Var}(M_i|T_{1i}=0,T_{2i}=1) > \hat{\omega}_1\widehat{Var}(M_i|T_{1i}=1,T_{2i}=0)$.

18 The finite-sample adjusted estimates are virtually identical, as should be expected given the sample size. For instance, the adjusted estimate of $\frac{\widehat{ACME}_2}{\widehat{ACME}_1}$ for the Disapproval dependent variable is 1.533, and the adjusted estimate of $\frac{\widehat{ACMET}_2}{\widehat{ACMET}_1}$ for the Punishment dependent variable is 1.796.

19 These results correspond to the case of "proportionate scaling up" presented in Table H2 in Appendix H.

20 For instance, the seminal experimental study by Tomz (2007) estimated audience cost effects between 16- and 32-percentage-point increases in disapproval in the context of security commitments and escalation management. Follow-up research in this area (e.g. Levendusky and Horowitz 2012) has also estimated effects of up to approximately 20 percentage points. Other experimental research on audience costs in areas of international legal and regulatory cooperation (e.g. Chaudoin 2014; Chilton 2015) have detected smaller effects of roughly 10-percentage-point increases in disapproval.

audience costs. In fact, additional evidence presented in Appendix G shows that another important mediation channel that contributes to these results is the fear of concrete international consequences or harm. In the scenario, this takes the form of concerns that other countries would follow suit in developing and potentially using similar weapons in the future, thus harming U.S. security in the long run.

In sum, legalization appears to have the potential to add to the domestic sources of credible commitment via multiple channels. However, the evidence presented here pertains to a specific international security context. Whether these findings would hold in other policy areas would be useful to explore in future research. For instance, in contexts where normative considerations are less salient, the morality channel may play a smaller role. The same argument could be made for the international consequences channel in contexts where the possibility of other countries reciprocating or retaliating is less of a concern. In such cases, would legalization continue to enhance audience costs, and if so, via what channels?

8 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 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 CCM context presented in this study.

Of course, the usefulness of these CCM methods is limited to experimental designs that feature multiple treatments, which are less common than single-treatment designs in many research settings. However, with the gradual accumulation of knowledge and empirical results in various academic subfields and program evaluation contexts, experimental research questions will increasingly evolve to require 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.

Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2019.31.

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 D. Snidal. 1998. “Why States Act Through Formal International Organizations.” *Journal of Conflict Resolution* 42(1):3–32.

Abbott, K. W., and D. Snidal. 2000. “Hard and Soft Law in International Governance.” *International Organization* 54(3):421–456.

Acharya, A., M. Blackwell, and M. Sen. 2016. “Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects.” *American Political Science Review* 110(3):512–529.

Albert, J. M. 2008. “Mediation Analysis via Potential Outcomes Models.” *Statistics in Medicine* 27(8):1282–1304.

Andrews, I., J. H. Stock, and L. Sun. 2019. “Weak Instruments in IV Regression: Theory and Practice.” *Annual Review of Economics*.

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton: Princeton University Press.

Arceneaux, K. 2008. “Can Partisan Cues Diminish Democratic Accountability?” *Political Behavior* 30(2):139–160.

Arceneaux, K., and R. Kolodny. 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):12–16.
Bansak, K. 2019. "Replication Materials for: Comparative Causal Mediation and Relaxing the Assumption of No Mediator-Outcome Confounding: An Application to International Law and Audience Costs." https://doi.org/10.7910/DVN/JLAOEN, Harvard Dataverse, V1.
Baron, R. M., and D. A. Kenny. 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., D. P. Green, and S. E. Ha. 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 S. E. Ha. 2011. "Mediation Analysis is Harder than it Looks." In *Cambridge Handbook of Experimental Political Science*, edited by J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia, 508–521. chapter 35, 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. 2014. "The Influence of International Human Rights Agreements on Public Opinion: An Experimental Study." *Chicago Journal of International Law* 15:110.
Chilton, A. S. 2015. "The Laws of War and Public Opinion: An Experimental Study." *Journal of Institutional and Theoretical Economics* 171(1):181–201.
Chilton, A. S., and M. Versteeg. 2016. "International Law, Constitutional Law, and Public Support for Torture." *Research & Politics* 3(1): 2053168016636413.
Cochran, W. G. 1963. *Sampling Techniques*. John Wiley & Sons.
Dai, X. 2005. "Why Comply? The Domestic Constituency Mechanism." *International Organization* 59(2):363–398.
Daniel, R. M., B. L. DeStavola, S. N. Cousens, and S. Vansteelandt. 2015. "Causal mediation analysis with multiple mediators." *Biometrics* 71(1):1–14.
Downs, G. W., D. M. Rocke, and P. N. Barsoom. 1996. "Is the Good News about Compliance Good News about Cooperation?" *International Organization* 50(3):379–406.
Fairchild, A. J., and H. L. McDaniel. 2017. "Best (but Oft-Forgotten) Practices: Mediation Analysis." *American Journal of Clinical Nutrition* 105(6):1259–1271.
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.
Flueck, J. A., and B. S. Holland. 1976. "Ratio Estimators and Some Inherent Problems in their Utilization." *Journal of Applied Meteorology* 15(6):535–543.
Franz, V. H. 2007 "*Ratios: A Short Guide to Confidence Limits and Proper Use*." arXiv:0710.2024 Technical report.
Fritz, M. S., and D. P. MacKinnon. 2007. "Required Sample Size to Detect the Mediated Effect." *Psychological Science* 18(3):233–239.
Gerber, A. S., and D. P. Green. 2012. "Mediation." In *Field Experiments: Design, Analysis, and Interpretation*, chapter 10, New York: W. W. Norton & Company.
Gleser, L. J., and J. T. Hwang. 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 E. A. Posner. 2005. *The Limits of International Law*. Oxford University Press.
Goldstein, J. 2001. *Legalization and World Politics*. MIT Press.
Goren, P., C. M. Federico, and M. C. Kittilson. 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.
Hong, G. 2015. *Causality in a Social World: Moderation, Mediation and Spill-Over*. John Wiley & Sons.
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., B. Jo, and E. A. Stuart. 2011a. "Commentary: Using Potential Outcomes to Understand Causal Mediation Analysis." *Multivariate Behavioral Research* 46(5):842–854.
Imai, K., L. Keele, and D. Tingley. 2010a. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15(4):309–334.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto. 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., L. Keele, and T. Yamamoto. 2010b. "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25:51–71.
Imai, K., D. Tingley, and T. Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176(1):5–51.
Imai, K., and T. Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21(2):141–171.
James, L. R., and J. M. Brett. 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.
Kam, C. D. 2005. "Who Toes the Party Line? Cues, Values, and Individual Differences." *Political Behavior* 27(2):163–182.
Koschat, M. A. 1987. "A Characterization of the Fieller Solution." *The Annals of Statistics* 15(1):462–468.
Kraemer, H. C., M. Kiernan, M. Essex, and D. J. Kupfer. 2008. "How and Why Criteria Defining Moderators and Mediators Differ Between the Baron & Kenny and Macarthur Approaches." *Health Psychology* 27(2S):S101.
Lehmann, E. L., and G. Casella. 2006. *Theory of Point Estimation*. Springer Science & Business Media.
Levendusky, M. S., and M. C. Horowitz. 2012. "When Backing Down is the Right Decision: Partisanship, New Information, and Audience Costs." *The Journal of Politics* 74(2):323–338.
Lipson, C. 1991. "Why are Some International Agreements Informal?" *International Organization* 45(4):495–538.
Loeys, T., W. Talloen, L. Goubert, B. Moerkerke, and S. Vansteelandt. 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 A. Smith. 2000. "Trust and Cooperation through Agent-Specific Punishments." *International Organization* 54(4):809–824.
Moravcsik, A. 2013. "Liberal Theories of International Law." In *Interdisciplinary Perspectives on International Law and International Relations*, edited by J. L. Dunoff and M. A. Pollack, 83–118. chapter 4, Cambridge: Cambridge University Press.
Morrow, J. D. 2000. "Alliances: Why Write Them Down?" *Annual Review of Political Science* 3(1):63–83.
Nelson, C. R., and R. Startz. 1990. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58(4):967–976.
Nicholson, S. P. 2012. "Polarizing Cues." *American Journal of Political Science* 56(1):52–66.
Olea, J. L. M., and C. Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics* 31(3):358–369.
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., S. C. Ropp, and K. Sikkink. 1999. *The Power of Human Rights: International Norms and Domestic Change*, vol. 66, Cambridge University Press.
Robins, J. M. 1997. "Causal Inference from Complex Longitudinal Data." In *Latent Variable Modeling and Applications to Causality*, 69–117. Springer.
Robins, J. M. 2003. "Semantics of Causal Dag Models and the Identification of Direct and Indirect Effects." In *Oxford Statistical Science Series*, 70–82.
Robins, J. M., and S. Greenland. 1992. "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology* 3(2):143–155.
Shpitser, I., and T. J. VanderWeele. 2011. "A Complete Graphical Criterion for the Adjustment Formula in Mediation Analysis." *The International Journal of Biostatistics* 7(1):1–24.
Simmons, B. A. 2009. *Mobilizing for Human Rights: International Law in Domestic Politics*. Cambridge University Press.
Simmons, B. A., and D. J. Hopkins. 2005. "The Constraining Power of International Treaties: Theory and Methods." *American Political Science Review* 99(4):623–631.
Slothuus, R., and C. H. de Vreese. 2010. "Political Parties, Motivated Reasoning, and Issue Framing Effects." *American Journal of Political Science* 72(3):630–645.
Staiger, D., and J. H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3):557–586.

Stock, J. H., J. H. Wright, and M. Yogo. 2002. "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business & Economic Statistics* 20(4):518–529.
Stock, J. H., and M. Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 80–108. Cambridge University Press.
Tchetgen, E. J. T., and I. Shpitser. 2012. "Semiparametric Theory for Causal Mediation Analysis: Efficiency Bounds, Multiple Robustness, and Sensitivity Analysis." *Annals of Statistics* 40(3):1816–1845.
Thoemmes, F., D. P. MacKinnon, and M. R. Reiser. 2010. "Power Analysis for Complex Mediational Designs Using Monte Carlo Methods." *Structural Equation Modeling* 17(3):510–534.
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. 2014. "A Unification of Mediation and Interaction: A 4-way Decomposition." *Epidemiology* 25(5):749–761.
VanderWeele, T. J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.
von Luxburg, U., and V. H. Franz. 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.
Withers, C. S. 1987. "Bias Reduction by Taylor Series." *Communications in Statistics-Theory and Methods* 16(8):2369–2383.
Zhang, Z. 2014. "Monte Carlo Based Statistical Power Analysis for Mediation Models: Methods and Software." *Behavior Research Methods* 46(4):1184–1198.