Assumption Smuggling in Intermediate Outcome Tests of Causal Mechanisms*

Matthew Blackwell† Ruofan Ma‡ Aleksei Opacic§

July 9, 2024

Abstract

Political scientists are increasingly attuned to the promises and pitfalls of establishing causal effects. But the vital question for many is not *if* a causal effect exists but *why* and *how* it exists. Even so, many researchers avoid causal mediation analyses due to the assumptions required, instead opting to explore causal mechanisms through what we call *intermediate outcome tests*. These tests use the same research design used to estimate the effect of treatment on the outcome to estimate the effect of the treatment on one or more mediators, with authors often concluding that evidence of the latter is evidence of a causal mechanism. We show in this paper that, without further assumptions, this can neither establish nor rule out the existence of a causal mechanism. Instead, such conclusions about the indirect effect of treatment rely on implicit and usually very strong assumptions that are often unmet. Thus, such causal mechanism tests, though very common in political science, should not be viewed as a free lunch but rather should be used judiciously, and researchers should explicitly state and defend the requisite assumptions.

*Working paper, comments welcome.

†Department of Government and Institute for Quantitative Social Science, Harvard University. web: http://www.mattblackwell.org, email: mblackwell@gov.harvard.edu

‡Department of Government, Harvard University. email: ruofan_ma@g.harvard.edu

§Department of Sociology, Harvard University. web: https://alekseiopacic.github.io, email: aopacic@g.harvard.edu

1 Introduction

Over the last few decades, causal inference has become a bedrock of quantitative research in political science, with greater care devoted to defining causal quantities of interest and stating and assessing the required assumptions. For good reasons, the main focus of this causal turn has been on establishing the existence (or nonexistence) of a causal effect on some interesting outcome. But scholars are often even more interested in *why* a causal effect exists as much as *if* a causal effect exists, leading to the proliferation of causal mechanism tests in empirical political science.

There has long been disagreement about the proper way to test for causal mechanisms, leading to a fractured landscape and difficulty in understanding how different tests relate. From a formal point of view, causal mechanisms have been most closely linked to mediation analysis, wherein the overall average treatment effect is decomposed into its direct (net of a mediating variable of interest) and indirect effects (Pearl, 2001; Imai, Keele and Yamamoto, 2010; Imai et al., 2011). Some scholars, however, object to causal mediation analysis because of the strong assumptions it requires (Green, Ha and Bullock, 2010; Acharya, Blackwell and Sen, 2016; Bullock and Green, 2021; Callis, Dunning and Tuñón, 2024). In particular, identification of the average direct and indirect effects relies on unverifiable assumptions that (i) the mediator is as-if randomized condition on pretreatment covariates, and (ii) there are no other post-treatment variables that confound the mediator-outcome relationship. This skepticism is reflected in the relative lack of causal mediation analyses in recent empirical political science articles. We analyzed 487 empirical papers published in the *American Political Science Review*, the *American Journal of Political Science*, and *The Journal of Politics* between 2022 and 2023 and found that 161 (33%) provided at least one quantitative test for causal mechanisms. As shown by Figure 1, however, of these 161 papers, only 16 (10%) employed a formal causal mediation analysis.

If these studies do not employ formal techniques of causal mediation, how do they assess causal mechanisms? Figure 1 shows it is common for researchers to employ an alternative approach to evaluating causal mechanisms, which in this paper we call *intermediate outcome tests* or IOTs. In these tests, researchers estimate the causal effect of the treatment on one or more potential mediators, often using the same research design used to estimate the main effect of the treatment on the outcome (Green, Ha

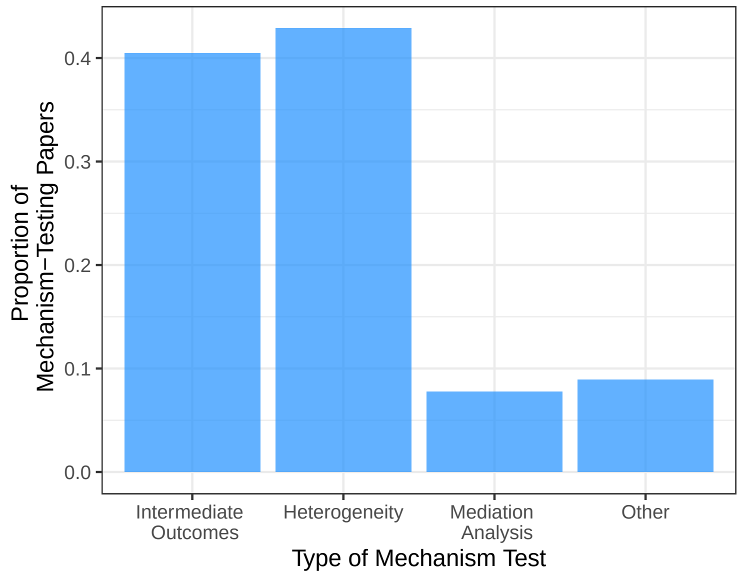

Figure 1: Proportion of papers analyzing causal mechanisms using different approaches from top journals in political science (APSR, AJPS, JOP), 2022–2023.

and Bullock, 2010; Callis, Dunning and Tuñón, 2024). Under this approach, if researchers detect an average treatment effect on a mediator, this mediator is considered part of a potential "mechanism" for the treatment effect on outcome. IOTs have clear, intuitive appeal since it seems to follow that, for M to mediate the relationship between treatment A and outcome Y, there must be an effect of A on M and then of M on Y. The (often implicit) argument for the IOT approach is that an average effect of A on M is a *necessary* and sometimes sufficient condition for M to be a causal mechanism. Thus, it is no surprise that this simple ad hoc approach is one of the dominant ways to estimate causal mechanisms, accounting for roughly 40% of the papers in our literature review.

In this paper, we show that, in an experimental context, intermediate outcome tests under randomization of treatment alone can neither establish nor rule out the existence of an indirect effect of treatment on the outcome through that mediator. Papers that imply otherwise are engaging (often unknowingly) in what we call *assumption smuggling* or presenting suspect conclusions that rely on strong hidden assumptions alongside well-grounded findings that depend on the more reasonable, explicitly stated assumptions. We establish the IOT's inability to inform mechanisms using sharp nonparametric bounds that contain all values of the indirect effect consistent with the observed data and the maintained assumptions. Under randomization of treatment alone, these bounds are never informative about the sign of the indirect effect for two reasons. First, the average effect of the treatment on the mediator might mask heterogeneous treatment effects that are important to the causal mechanism. Second, establishing indirect effects requires strong evidence or assumptions about the effect of the mediator on the outcome at the individual level.

To help researchers navigate this, we further introduce a set of assumptions that, though very rarely discussed in practice, do enable the IOT procedure to have a meaningful interpretation. Specifically, we investigate a *monotonic mediator response* assumption that the units' value of the mediator can only be affected by treatment in one direction. Under this assumption, we show that a lack of treatment effect on the mediator can, in fact, rule out indirect effects through that mediator. We also derive the nonparametric sharp bounds for the average indirect effect under this assumption, showing that IOTs can "approximately" rule out mechanisms in that small (zero) effects of the treatment on the outcome imply small (zero) indirect effects. Of course, monotonicity is a strong, untestable assumption that will frequently be violated in applied practice, and there is no design-based way to ensure it holds. Thus, it is critical for researchers to state this assumption clearly and to interrogate it if they wish to use IOTs to assess causal mechanisms at all.

Can IOTs ever establish the sign of an indirect effect? We also show that, unfortunately, it is nearly impossible to establish the sign of an indirect effect even under very strong assumptions. In particular, we show that even if we *assume* a positive average effect of the mediator on the outcome, the sharp nonparametric bounds on the indirect effect will never be informative about the sign of that effect. Thus, while the monotonic mediator response assumption may allow researchers to rule out potential pathways, no intermediate outcome test can establish the simple *existence* of a causal mechanism even if we make heroic assumptions about the mediator-outcome relationship.

The contributions of this paper are twofold. First, we highlight for applied researchers that this ad hoc, widely used approach to analyze causal mechanisms actually relies on strong and unusually unstated assumptions; relaxing the assumptions, as we show, can render substantive conclusions invalid. Second, by clarifying the assumptions needed, we provide guidance for applied researchers looking to provide substantive meaning for the interpretation of these IOTs. The takeaway message of this paper reinforces what has been argued elsewhere: establishing causal mechanisms is difficult and requires strong assumptions regardless of the approach. Researchers may not avoid this difficulty even when using IOTs.

Our work contributes to a broader literature on causal mechanisms and (in)direct effects in multiple fields. Many scholars have formalized the concept of direct and indirect effects in terms of potential outcomes and have derived assumptions that can point identify these quantities (Pearl, 2001; Imai, Keele and Yamamoto, 2010). Most similar to our own work is Imai, Tingley and Yamamoto (2013), which derived nonparametric bounds for indirect effects under several different experimental designs, and Glynn (2012), which showed how individual-level heterogeneity in treatment effects renders many indirect effect tests ineffective. We expand on their results by showing how different causal assumptions can change the bounds for a single experiment. Several studies have proposed alternative ways to assess causal mechanisms, such as implicit mediation (Bullock and Green, 2021) and effect heterogeneity (Fu and Slough, 2024). However, these papers show that these alternatives also rely on strong assumptions. An alternative definition of causal mechanisms centers on a quantity known as the "controlled direct effect" (CDE), which captures the average direct effect of a treatment variable while holding the mediator fixed at a given value (Goetgeluk, Vansteelandt and Goetghebeur, 2008; Acharya, Blackwell and Sen, 2016, 2018; Zhou and Wodtke, 2019; Blackwell and Strezhnev, 2022). While identifying the CDE requires weaker assumptions than mediation, many mediation-skeptical scholars object to these approaches because they require us to identify the effect of the mediator on the outcome.[1]

This paper proceeds as follows. In Section 2, we discuss the setting of our two empirical applications, showing how researchers use IOTs in practice. Since the phrase "causal mechanisms" means

[1]Some recent work investigates how to handle multiple mediating variables (Imai and Yamamoto, 2013; VanderWeele, Vansteelandt and Robins, 2014; Daniel et al., 2015; Zhou and Yamamoto, 2023). These approaches typically rely on stronger identification assumptions than in the single-mediator case, and multiple mechanism tests have yet to be widely adopted in political science. For these reasons, we focus on the single-mediator setting.

different things to different scholars, we next lay out our preferred notation and definitions for causal effects and causal mechanisms in Section 3. In Section 4, we formally analyze intermediate outcomes tests using a principal stratification approach and derive sharp nonparametric bounds under different assumptions. Section 5 applies these bounds to two recent studies, showing the limited amount that can be learned about indirect effects with these tests. Section 6 discusses how scholars might consider mechanisms when they cannot meet the assumptions we discuss in this paper. Section 7 concludes and describes avenues for future research.

2 Two Motivating Examples

2.1 Reducing Outgroup Prejudice

An important literature in political psychology examines intergroup prejudices and which forms of individual-oriented interventions might durably reduce such prejudices. An approach that has garnered empirical support in a growing literature is the use of interpersonal conversations (Galinsky and Moskowitz, 2000; Bruneau and Saxe, 2012; Broockman and Kalla, 2016; Adida, Lo and Platas, 2018; Simonovits, Kezdi and Kardos, 2018; Audette, Horowitz and Michelitch, 2020; Lowe, 2021; Williamson et al., 2021), but an unresolved question in this recent work is which "narrative strategies" are most effective at reducing prejudice. For example, an interpersonal conversation that emphasizes taking the perspective of an outgroup member ("perspective taking") may yield different outcomes from one that simply describes the negative experiences of the outgroup (" perspective getting").

To evaluate the effects of distinct forms of interpersonal conversation on reducing prejudice, Kalla and Broockman (2023) implement three field experiments alongside a survey experiment. In the field experiments, volunteer canvassers went door to door (or phoned houses, in the case of the third) to a sample of registered voters across multiple states, engaging in conversation about unauthorized immigrants. Each respondent was randomly assigned to one of several narrative strategies. In the survey experiment, researchers showed respondents a picture of an outgroup member and asked them to engage in an exercise corresponding to different narrative strategies.

Across these experiments, the authors find narratives that simply describe the outgroup's experience of discrimination consistently and durably reduce prejudicial attitudes as much as the narratives that additionally ask respondents to imagine how they may feel if they had a similar experience. But why does perspective-getting reduce prejudice? Addressing this question is important not only because of its theoretical dividends but also because it offers important insights for practitioners. Given that prejudice interventions are often expensive and unscalable, discerning *how* perspective-getting reduces exclusionary attitudes facilitates the development of more cost-effective interventions that target these mechanisms. The authors implement an IOT to investigate mechanisms, showing positive, statistically significant effects of perspective-getting narratives on a handful of potential mediators, and argue that intervention "activates multiple mechanisms that have been found to reduce prejudice" (Kalla and Broockman, 2023, p. 188).

2.2 Transitional Justice Museums and Support for Democracy

Transitional justice policies are frequently credited with facilitating democratic consolidation, reconciliation, and peace-building in post-conflict societies (De Brito, Enríquez and Aguilar, 2001; Horne, 2014; Nalepa, 2010). Less is known, however, about how such policies operate at the individual level and through what mechanisms. In a field experiment in Chile, Balcells, Palanza and Voytas (2022) study one particular instance of transitional justice policy: museums. Students were randomized into either a control condition or a treatment condition consisting of an hour-long tour of a museum memorializing victims of General Augusto Pinochet's dictatorship. The authors then fielded a survey about beliefs about democracy and found that treated students "display greater support for democratic institutions, are more likely to reject institutions associated with the repressive period, and are more supportive of restorative transitional justice policies" (Balcells, Palanza and Voytas, 2022, p. 496).

Identifying how such an intervention increases support for democratic institutions could help inform the design of other, potentially less costly, policies. The authors implement an IOT procedure using a battery of emotion-oriented questions to understand the mechanisms underlying their estimated effects. They find that treated students demonstrate higher levels of emotion in both a positive (e.g., feeling stimulated, inspired, interested) and negative (e.g., feeling tense, nervous, embarrassed) sense. The authors take their IOT results as evidence that the "emotional experience" elicited by a museum visit is an important pathway that explains its promotion of democratic sentiment.

3 Causal mechanisms: notation and quantities of interest

These two studies illustrate how many political scientists generate substantive conclusions about "causal mechanisms," the meaning of which is the source of some debate in the methodological and statistical literature. We explore the interpretation of mediation-style indirect effects as one measure of a causal mechanism. To explore this, we focus on a simple yet common empirical setting in the social sciences: the effect of a binary treatment on a binary outcome with a binary mediator. Obviously, this does not capture all settings in political science, but many of the core ideas we present generalize to other settings easily. In this section, we introduce basic notation and some well-understood issues in the identification of direct and indirect effects; this sets the stage for the paper's contributions in the following section, in which we formalize the assumptions necessary for the IOT approach to be informative about the mechanism.

Let $A_i \in \{0, 1\}$ be a binary treatment variable, $M_i \in \{0, 1\}$ be the binary mediator, and $Y_i \in \{0, 1\}$ be the binary outcome. In some cases, there may also be a vector of pretreatment covariates, $\mathbf{X}_i$, to aid identification or make estimates more efficient. We refer to the $(A_i, M_i, Y_i, \mathbf{X}_i)$ as the observed data and assume the vector is an independent and identically distributed draw from some population distribution.

We use the potential outcomes framework for causality. Let $Y_i(a, m)$ be the outcome that unit i would have if (potentially counter to fact) they received treatment $A_i = a$ and mediator $M_i = m$. We make the usual consistency assumption that ties these potential outcomes to observed outcomes:

$$Y_i = Y_i(a, m) \qquad \text{if} \quad A_i = a, M_i = m.$$

In mediation analyses, we are also concerned about the effect of treatment on the mediator, so we define potential outcomes for the mediator, $M_i(a)$ as well. In this framework, causal effects are contrasts between different potential outcomes.

With this notation, it is possible to define potential outcomes that involve "natural" values of the mediator, such as $Y_i(1, M_i(0))$, which is the outcome that unit i would take if it was assigned to treatment but had received the mediator value under the control condition. These types of potential outcomes are important to mediation, but they create challenges in estimation and interpretation due to their "cross-world" nature. In particular, a unit cannot simultaneously be subject to treatment and control conditions at the same time, so $Y_i(1, M_i(0))$ is not observable even in principle.

The first causal question in this setting is generally "what is the overall effect of treatment?" To this end, the potential outcome is just setting the treatment as $Y_i(a) = Y_i(a, M_i(a))$, and we define the average treatment effect (ATE) as

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)].$$

Similarly, we define the average treatment effect on the mediator (ATM) as

$$\alpha = \mathbb{E}[M_i(1) - M_i(0)].$$

Finally, the average natural indirect effect (ANIE) for a given treatment level is defined as

$$\delta(a) = \mathbb{E}[Y_i(a, M_i(1)) - Y_i(a, M_i(0))]. \tag{1}$$

The ANIE captures the "path-specific" idea of causal mechanisms since it is the effect of the mediator induced by a change in treatment status, holding the direct effect of treatment constant. In a similar spirit, we define the average natural direct effect (ANDE) as

$$\zeta(a) = \mathbb{E}[Y_i(1, M_i(a)) - Y_i(0, M_i(a))], \tag{2}$$

which is the direct effect of the treatment holding the mediator to its natural value under treatment level a. An attractive property of these mediation quantities is that they decompose the ATE,

$$\tau = \delta(a) + \zeta(1 - a), \tag{3}$$

allowing a researcher to determine how much of an overall effect is due to the mediator or other factors (Pearl, 2001; Imai, Keele and Yamamoto, 2010).

3.1 Identifying assumptions

The identification of the above causal effects from observational data requires assumptions. In this paper, we assume a setting where researchers believe the main effects of treatment are identified but are more skeptical of assumptions about the relationship between the mediator and the outcome. We begin with the core identifying assumption of an experimental study.

Assumption 1 (Randomization)**.** *For all* a, a', m, $\{Y_i(a, m), M_i(a')\} \perp\!\!\!\perp A_i \mid \mathbf{X}_i$.

Assumption 1 is justified by randomized experiments but is also the key identifying assumption of papers that invoke selection on observables to justify causal interpretations of estimates. While the latter setting is often thought to be less plausible, a large body of work focuses on so-called "natural" experiments that could make this assumption believable. Under this assumption, we can identify the ATE as the average of covariate-specific effects,

$$\tau = \mathbb{E}\left[\mathbb{E}[Y_i \mid A_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i \mid A_i = 0, \mathbf{X}_i]\right].$$

This randomization assumption encodes our belief in the basic design of an empirical paper. However, identifying the direct and indirect effects relies on stronger assumptions about the relationship between the mediator and the outcome. In particular, Pearl (2001) and Imai, Keele and Yamamoto (2010) have shown that we can identify the ANIE and ANDE by assuming that the mediator is also as-if randomized conditional on treatment and the covariates.

Assumption 2 (Mediator as-if randomization)**.** $Y_i(a', m) \perp\!\!\!\perp M_i(a) \mid A_i = a, \mathbf{X}_i$

Many authors have expressed skepticism about Assumption 2 since it requires that there are no (measured *or* unmeasured) post-treatment confounders for the relationship between the mediator and the outcome. This means the effect of M_i on Y_i must be identified using the exact same design as the overall effect of treatment.[2] This condition is often difficult to sustain when the mediator is

[2] One alternative approach that has garnered some attention in recent years involves estimating a quantity known as the controlled direct effect (CDE), which captures the causal effect of a treatment when the mediator is fixed to a given value. This quantity is attractive because it is identified under a weaker assumption than Assumption 2, one that allows for the existence of observed "intermediate confounders" (confounders that are affected by the treatment and which affect the mediator and outcome). A drawback of this approach is that it does not quantify the extent of mediation through M_i and only enables a researcher to rule out the existence of alternative mediators other than M_i (see Acharya, Blackwell and Sen (2016)).

observed at its natural value after treatment, as is commonly the case. Indeed, while many are willing to believe the identifying assumptions for treatment, assuming the same condition for the mediator is much less plausible.

3.2 Assumption smuggling

The rest of this paper will assume the position of the mediation skeptic who is willing to believe Assumption 1 but not Assumption 2. We will show what these mediation-skeptical researchers can learn about causal mechanisms and indirect effects from intermediate outcome tests alone.

We define *assumption smuggling* to be the (perhaps unintentional) practice of presenting evidence for a claim that only tests the claim under additional assumptions that are strong and undisclosed. (We limit this definition to strong assumptions to differentiate this practice from omitting regularity conditions like the existence of fourth moments or other technicalities.) The disclosure requirement for an assumption may be satisfied by common knowledge of an assumption (a linear functional form, perhaps) or by explicit disclosure when presenting evidence for the claim.

Empirical studies often contain a section on "mechanisms" with empirical tests that purport to elucidate why a causal effect exists without laying out the assumptions these tests require. This is a prime example of assumption smuggling: attempting to maintain nominally weak assumptions with empirical tests that actually require a stronger set of unmentioned assumptions. Of course, like with unattended luggage in an airport, most instances of assumption smuggling are unintentional. Many strong assumptions have been embedded in the parametric models commonly used in political science, and many of us have built our statistical intuitions from that foundation. But one benefit of the recent "credibility revolution" has been to encourage researchers to be more clear about the assumptions of their research designs. Our goal is to help mediation-skeptic scholars who wish to test causal mechanisms do so with a clear articulation of what exactly they are assuming.

4 Intermediate Outcome Tests

We now introduce the first common test for causal mechanisms in the empirical literature: the intermediate outcome test. The logic of an IOT is quite compelling. We assume that a researcher believes Assumption 1 to justify interpreting the estimated effect of A_i on Y_i as causal. That same assumption will also justify estimating the causal effect of treatment on the mediator with the same basic strategy:

$$\alpha = \mathbb{E}[M_i(1) - M_i(0)] = \mathbb{E}\left[\mathbb{E}[M_i \mid A_i = 1, \mathbf{X}_i] - \mathbb{E}[M_i \mid A_i = 0, \mathbf{X}_i]\right].$$

Scholars using IOT usually present comparisons of τ and α as a "necessary but not sufficient argument" for causal mechanisms. Suppose, for example, we have estimated a positive ATE, $\widehat{\tau} > 0$. We might believe that a certain mediator captures the mechanism of interest in the sense that a positive indirect effect through it explains the positive $\widehat{\tau}$. If indirect effects are effects along the path $A_i \longrightarrow M_i \longrightarrow Y_i$, then we might believe that a positive average treatment effect on the mediator would establish the first of these arrows. Under this argument, a positive $\widehat{\alpha}$ is necessary but not sufficient since we also require information about the $M_i \longrightarrow Y_i$ arrow. Green, Ha and Bullock (2010, p. 207) explain the logic of this approach (see Callis, Dunning and Tuñón (2024, pp. 61–2) for a similar description):

> A more judicious approach at this juncture in the development of social science would be to encourage researchers to measure as many outcomes as possible when conducting experiments... With many mediators and only one intervention, this kind of experiment cannot identify which of the many causal pathways transmit the effect of the treatment, but if certain pathways are unaffected by the treatment, one may begin to argue that they do not explain why [the intervention] works.

In other words, under this argument, we might view IOTs as a kind of falsification test—we hope they have the potential to rule out mechanisms such that passing the test has some informative value. Scholars often take this strategy explicitly because they are uncomfortable with the assumptions required to estimate the causal relationship between the mediator and the outcome.

Unfortunately, the "necessary but not sufficient" argument above has been shown to be invalid. Glynn (2012) showed that the product of the average effect of A_i on M_i and the average effect of M_i on Y_i does not produce estimates of indirect effects when treatment effects can vary from unit to unit.[3] We extend those results by providing nonparametric bounds on what we can learn about the indirect effect from the maintained assumptions of the mediation skeptic.

4.1 Principal stratification approach to mediation

One advantage of the binary setting we explore in this paper is that it allows us to follow a principal stratification approach (see also, Bullock and Green, 2021). In particular, we can stratify all units into four groups based on how their mediator value responds to treatment:

- mechanism compliers, $M_i(0) = 0, M_i(1) = 1$,
- mechanism defiers, $M_i(0) = 1, M_i(1) = 0$,
- always takers $M_i(0) = M_i(1) = 1$,
- never takers $M_i(0) = M_i(1) = 0$.

We borrow the terminology of compliers and defiers from the literature on instrumental variables to mean units that comply with or defy the proposed causal mechanism (rather than treatment assignment as in the instrumental variables setting). We can define an individual-level indirect effect as

$$\delta_i(a) = Y_i(a, M_i(1)) - Y_i(a, M_i(0)),$$

which makes it clear that the only individuals who will have an indirect effect are either mechanism compliers or defiers. That is $\delta_i(a) = 0$ for the always and never takers. This fact motivates the IOT

[3] Indeed, Green, Ha and Bullock (2010) recognize this when they say, "this kind of analysis makes some important assumptions about homogeneous treatment effects, but the point is that this type of exploratory investigation may provide some useful clues to guide further experimental investigation." We have not found researchers using IOTs to either acknowledge these assumptions or be as cautious in the interpretation of IOTs as Green, Ha and Bullock (2010) recommend.

approach: if no effect of A_i on M_i implies no indirect effect at the individual level, it feels natural that it should apply at the average level. Unfortunately, this is not the case.

To see why the IOT approach falls down without further assumptions, let us relate the average treatment effect on the mediator (ATM) to these principal strata. We can define the size of each of these groups as

$$\rho_{st} \triangleq \Pr[M_i(1) = s, M_i(0) = t].$$

With these, we can show that ATM is equal to the proportion of mechanism compliers minus the proportion of mechanism defiers,

$$\alpha = \mathbb{E}[M_i(1) - M_i(0)] = 1 \cdot \rho_{10} + (-1) \cdot \rho_{01} + 0 \cdot \rho_{11} + 0 \cdot \rho 00 = \rho_{10} - \rho_{01},$$

where the second equality follows from an application of the iterated expectations. The always-takers and never-takers contribute nothing because their values of M_i never change with treatment.

We can also write the ANIE in terms of these principal strata,

$$\begin{aligned}
\delta(a) &= \mathbb{E}[Y_i(a, M_i(1)) - Y_i(a, M_i(0))] \\
&= \sum_{s=0}^{1} \sum_{t=0}^{1} \mathbb{E}[Y_i(a, s) - Y_i(a, t) | M_i(1) = s, M_i(0) = t] \cdot \rho_{st} \\
&= \mathbb{E}[Y_i(a, 1) - Y_i(a, 0) | M_i(1) = 1, M_i(0) = 0] \cdot \rho_{10} \\
&\qquad - \mathbb{E}[Y_i(a, 1) - Y_i(a, 0) | M_i(1) = 0, M_i(0) = 1] \cdot \rho_{01}.
\end{aligned} \tag{4}$$

The last expression shows that the ANIE is a function of the average effect of the mediator on the outcome for the mechanism compliers and defiers. The compliers contribute positive effects of the mediator while defiers contribute their negative effects.

The IOT approach restricts itself from identifying the ATM, or $\rho_{10} - \rho_{01}$ in the principal strata framework. With no further assumptions, what can that tell us about mechanisms? From (4), we can see that we can have a non-zero ANIE even with an ATM of exactly zero. In particular, the ATM could be zero if there are equal numbers of mechanism compliers and defiers—that is, if an equal number of respondents' mediators are positively and negatively affected by treatment—but the effect of M_i on Y_i is different for these groups. Thus, IOTs alone cannot rule out an indirect effect without further assumptions, leaving their interpretation as falsification tests suspect. Furthermore, non-zero ATMs are also *not sufficient* for indirect effects because any difference between ρ_{10} and ρ_{01} in (4) can be offset by differences in the effect of M_i on Y_i in each of these groups. Thus, when we maintain the set of mediation-skeptical assumptions, intermediate outcomes provide almost no information about causal mechanisms, at least in the form of indirect effects.

None of these results depend on any uncertainty in our estimates. These results show what we could learn about indirect effects if we had access to infinite data and thus no sampling uncertainty. In reality, we will estimate these quantities, leading these tests to have even less informativeness.

4.2 The hidden assumptions of intermediate outcome tests

If intermediate outcomes are largely uninformative for mediation skeptics, why do we see them in widespread use? In this section, we show how to recover the intended informativeness of IOTs by making additional assumptions. We call these the *hidden assumptions* since they are often implicitly invoked when using IOTs.

We begin with assumptions on the relationship between the treatment and moderator since most researchers will be more comfortable with assumptions on the primary research design than on the effects of the mediator. We can recover part of the informativeness of the IOT approach by assuming that the treatment only affects the mediator in one direction.

Assumption 3 (Monotonic Mediator Response (MMR))**.** $\mathbb{P}[M_i(1) \geq M_i(0)] = 1$.

In words, Assumption 3 states that an individual's potential mediator value is always weakly greater if they were to receive the treatment. This rules out the existence of mediator-defiers in the population (that is, units for which $M_i(1) = 0, M_i(0) = 1$), meaning that $\rho_{01} = 0$ and the ATM will be equal to ρ_{10}. Applying this assumption to (4) allows us to write the ANIE as

$$\delta(a) = \mathbb{E}[Y_i(a, 1) - Y_i(a, 0) | M_i(1) = 1, M_i(0) = 0] \cdot \rho_{10}, \tag{5}$$

i.e., a scaled (by ρ_{10}) controlled direct effect (CDE) of M_i on Y_i among the mechanism compliers. Thus, under MMR, we recover a weak "product rule" interpretation of the indirect effect. We can write the

ANIE as the product of the effect of the treatment on the mediator and the effect of the mediator on the outcome among units whose mediator is affected by treatment, or

$$\delta(a) = \mathbb{E}[Y(a,1) - Y(a,0) | M(1) = 1, M(0) = 0] \cdot \text{ATM}. \quad (6)$$

Thus, under MMR, an average effect of treatment on the moderator of zero rules out an average indirect effect, and the IOT can now properly function as a falsification test. Unfortunately, intermediate outcome tests remain insufficient for establishing causal mediation because even if the ATM is non-zero, the effect of M_i on Y_i for the mechanism compliers may still be zero. Is the MMR assumption plausible? The answer to this question will, of course, depend on the context. There may be many settings where this is a perfectly reasonable assumption. A survey experiment with a prompt designed to anger respondents is unlikely to make some respondents less angry. The difficulty with this assumption is that it must be true for all units to make the above interpretations correct, which may be difficult to sustain in many cases.

4.3 Sharp bounds for the indirect effect

The previous section showed when IOTs can be used to rule out causal mechanisms. These results, however, are somewhat broad, and it would be useful to have a more precise sense of what we can learn about the indirect effect from a given empirical setting. We next turn to establishing nonparametric bounds for the indirect effect, which are the values of an indirect effect that are consistent with a set of empirical results and assumptions (Manski, 1995). To do so, we use the linear programming approach of Balke and Pearl (1997) and Sachs et al. (2023).[4]

Let $p_{ym \cdot a} = \mathbb{P}(Y_i = y, M_i = m \mid A_i = a)$ be the joint distribution of the outcome and the mediator within levels of treatment. Imai, Tingley and Yamamoto (2013) derived the sharp nonparametric

[4]We ignore covariates in these derivations for presentational simplicity, but it is possible to incorporate covariates for either identification or for narrowing the bounds (Levis et al., 2023).

bounds for the ANIE under the mediation-skeptic assumptions (that is, Assumption 1) as

$$\max \left\{ \begin{array}{l} -p_{10\cdot 0} - p_{11\cdot 0} \\ -p_{01\cdot 1} - p_{11\cdot 1} - p_{11\cdot 0} \\ -p_{00\cdot 1} - p_{10\cdot 1} - p_{10\cdot 0} \end{array} \right\} \leqslant \bar{\delta}(0) \leqslant \min \left\{ \begin{array}{l} p_{00\cdot 0} + p_{01\cdot 0} \\ p_{01\cdot 1} + p_{11\cdot 1} + p_{01\cdot 0} \\ p_{00\cdot 0} + p_{00\cdot 1} + p_{10\cdot 1} \end{array} \right\},$$

$$\max \left\{ \begin{array}{l} -p_{00\cdot 1} - p_{01\cdot 1} \\ -p_{01\cdot 1} - p_{01\cdot 0} - p_{11\cdot 0} \\ -p_{00\cdot 0} - p_{00\cdot 1} - p_{10\cdot 0} \end{array} \right\} \leqslant \bar{\delta}(1) \leqslant \min \left\{ \begin{array}{l} p_{10\cdot 1} + p_{11\cdot 1} \\ p_{01\cdot 0} + p_{11\cdot 0} + p_{11\cdot 1} \\ p_{00\cdot 0} + p_{10\cdot 0} + p_{10\cdot 1} \end{array} \right\}.$$

These bounds tell us the most we can learn about the ANIE from the joint distribution of the data (including the mediator) under the mediation-skeptical assumptions. Note that because $p_{ym\cdot a}$ is non-negative, zero will always be contained in these bounds, which is consistent with the earlier discussion that IOTs are uninformative about indirect effects under the mediation-skeptical assumptions. With these bounds, we cannot rule out a mechanism even if the ATM is zero since non-zero indirect effects will always be in the bounds, no matter the values of the ATM. Thus, we are extremely limited in what we can learn about indirect effects from IOTs.

We can derive sharp bounds for the indirect effect under different assumptions. The bounds under MMR provide bounds that we can write in terms of the ATM.

Proposition 1. *Under Assumption 3, the sharp bounds for the ANIE are*

$$\max \{-ATM, -p_{10\cdot 0}\} \leq \delta(0) \leq \min \{ATM, p_{00\cdot 0}\}, \tag{7}$$

$$\max \{-ATM, -p_{01\cdot 1}\} \leq \delta(1) \leq \min \{ATM, p_{11\cdot 1}\} \tag{8}$$

These sharp bounds under the MMR assumption provide a more nuanced interpretation of IOTs. When the ATM is relatively small (compared to $p_{10\cdot 0}$, $p_{00\cdot 0}$, $p_{01\cdot 1}$, and $p_{11\cdot 1}$), the sharp bounds for the ANIE will be $[-\text{ATM}, \text{ATM}]$. Thus, when the ATM is close to 0, the bounds will also be close to zero, and we will be able to rule out large indirect effects. At the extreme, if the ATM is zero, then the ANIE is also point identified as zero, as the discussion above implies. This result implies that IOTs under MMR are "approximately necessary" in that small (zero) ATMs imply small (zero) indirect effects. Again, however, we emphasize that this interpretation of an IOT relies on the monotonic mediator response assumption, which cannot be guaranteed to hold by design.

What assumptions would be necessary to allow an IOT to establish the presence of an indirect effect? Any such assumption would need to place restrictions on the causal relationship between the mediator and the outcome. Indeed, as shown by Pearl (2001) and Imai, Keele and Yamamoto (2010), we must assume that the mediator is as-if randomized with respect to the outcome (conditional on pretreatment covariates) to point identify the ANIE. Of course, the mediation-skeptical approach seeks to avoid any such assumptions.

One potential justification for the IOT approach is if past empirical evidence or theory implies there should be an average effect of the mediator on the outcome. If the treatment "activates" the pathway from A_i to M_i in the experimental data (that is, there is an ATM), then this should provide sufficient evidence to establish the causal mechanism if the supposition is true, or so the reasoning goes. In the next proposition, we provide sharp bounds on the indirect effect under MMR and knowledge that the average treatment effect of M_i on Y_i is positive. Even under these very strong assumptions, the bounds always include an indirect effect of zero.

Proposition 2. *Suppose that Assumption 3 holds and* $\mathbb{E}[Y_i(1,1) - Y_i(1,0)] \geq 0$. *Then,*

$$\max \left\{ \begin{array}{l} -ATM \\ p_{10\cdot 1} - p_{10\cdot 0} - p_{00\cdot 0} \\ -p_{11\cdot 1} - p_{00\cdot 1} - p_{10\cdot 0} \\ -p_{01\cdot 1} \end{array} \right\} \leq \bar{\delta}(1) \leq \min \left\{ \begin{array}{l} ATM \\ p_{11\cdot 1} + p_{10\cdot 0} + p_{00\cdot 0} \\ 2p_{11\cdot 1} + p_{00\cdot 1} + p_{00\cdot 1} \\ p_{11\cdot 1} \end{array} \right\}, \tag{9}$$

and these bounds are sharp. Furthermore, all lower bounds are nonpositive, and all upper bounds are nonnegative.

Proposition 2 shows that even if we assume monotonic mediator response *and* a positive average treatment effect of the mediator on the outcome, we still cannot learn the direction of the indirect effect. Equation 6 shows why this is the case—the indirect effect focuses on the effect of M_i on Y_i for the mechanism compliers, whereas the average treatment effect of M_i on Y_i is the effect averaged across all units. Thus, even under these assumptions, IOTs cannot be sufficient tests for causal mechanisms.

4.4 Estimation and inference for the bounds

The above bounds are all functions of the population probabilities $p_{ym \cdot a}$, but researchers only have access to the observed data $\{(A_i, M_i, Y_i, \mathbf{X}_i)\}_{i=1,\ldots,n}$. To perform estimation and inference for these bounds, we rely on the intersection bounds approach of Chernozhukov, Lee and Rosen (2013). Traditional plug-in estimators for the bounds tend to be biased so that the plug-in bounds are narrower than the population bounds because of the use of the minimum and maximum operators. Furthermore, the estimated standard error of the selected bounding expression would ignore the uncertainty that the other bounding expressions also have, which contributes to uncertainty about the overall bound. The approach of Chernozhukov, Lee and Rosen (2013) uses precision-corrected versions of each bounding expression to obtain (a) upper and lower bound estimates that are half-median unbiased and (b) provides confidence intervals for the true indirect effect that cover the true value of the indirect effect at nominal rates. Half-median unbiased means that the probability that the estimated upper (lower) bound is greater (less) than the population upper (lower) bound is at least one-half. This is a common desirable property of estimators in partially identified settings like ours where previous work has shown that traditional unbiasedness is not possible to achieve (Hirano and Porter, 2012). In the next section, we implement these empirical procedures for our two motivating examples.

5 Empirical Analyses

5.1 Reducing Outgroup Prejudice

We first re-analyze data from the Kalla and Broockman (2023) study of intergroup prejudice. In one of their studies, the authors showed respondents a picture of an outgroup member (either undocumented immigrants or transgender people) and then asked them to engage in a randomly assigned exercise corresponding to different narrative strategies.[5] We focus on the pooled (i.e., across the unauthorized immigrants and transgender people arms) effects of the "perspective getting with essay" treatment relative to the baseline condition, since this treatment yields the highest effect on prejudice

[5] These are traditional perspective-taking, analogic perspective-taking, perspective getting with an essay task to summarize the story heard, and perspective getting without the essay task.

reduction of the four treatment conditions considered. In this treatment condition, respondents read a story about an outgroup member and were then asked to summarize the story they just read.

To assess which mechanisms explain these effects, the authors asked respondents four additional questions as potential mediators: (i) whether respondents have a lot in common with the outgroup; (ii) whether the suffering of the outgroup is of concern ; (iii) whether outgroup members face challenges that are no fault of their own, and (iv) whether the outgroup's name elicits thoughts about individual people or a group of people. Importantly for our purposes, the authors implement an IOT approach on these four mediators.

Since our sharp bounds for the indirect effect (as described in Section 4.3) rely on dichotomous mediators and outcomes, we first dichotomize the four mediators and the outcome variable at their medians. We then implement the IOT approach, estimating IOT effects without covariate controls. We do so because the treatment is randomly assigned in this study. As shown in the right panel of Figure 2, these IOT estimates largely replicate the direction, magnitude, and statistical significance of those reported by Kalla and Broockman (2023): perspective-getting with an essay task has a positive and statistically significant effect on whether the suffering of the outgroup is of concern (0.04, 95% confidence interval of [0.03, 0.06]) and whether respondents have a lot in common with the outgroup (also 0.04, 95% confidence interval of [0.03, 0.06]).[6] The authors rule out the last of these mediators because placebo interventions also induced a sense of commonality with the outgroup. Thus, they argue that concern about outgroup suffering is likely a key pathway for the effect of their intervention.[7]

Does this conclusion hold if we implement our nonparametric bounds for the indirect effect? Figure 2, right panel, suggests it does not. This panel displays the sharp bounds for the indirect effect with the treatment reference category of $A = 0$, i.e., $\delta(0)$, and calibrates these bounds both against zero (dotted black line) and the overall ATE of the intervention (dotted red line), which was 0.03.

[6] Note that the authors also find a positive effect of perspective-getting with an essay task on whether outgroup members face challenges that are no fault of their own, a finding that does not replicate without controlling for the battery of covariates included in the authors' models.

[7] Of course, these different mediators overlap to some degree. We treat each mediator separately, but a fruitful path for future research would be to incorporate multiple mediators into these bounds.

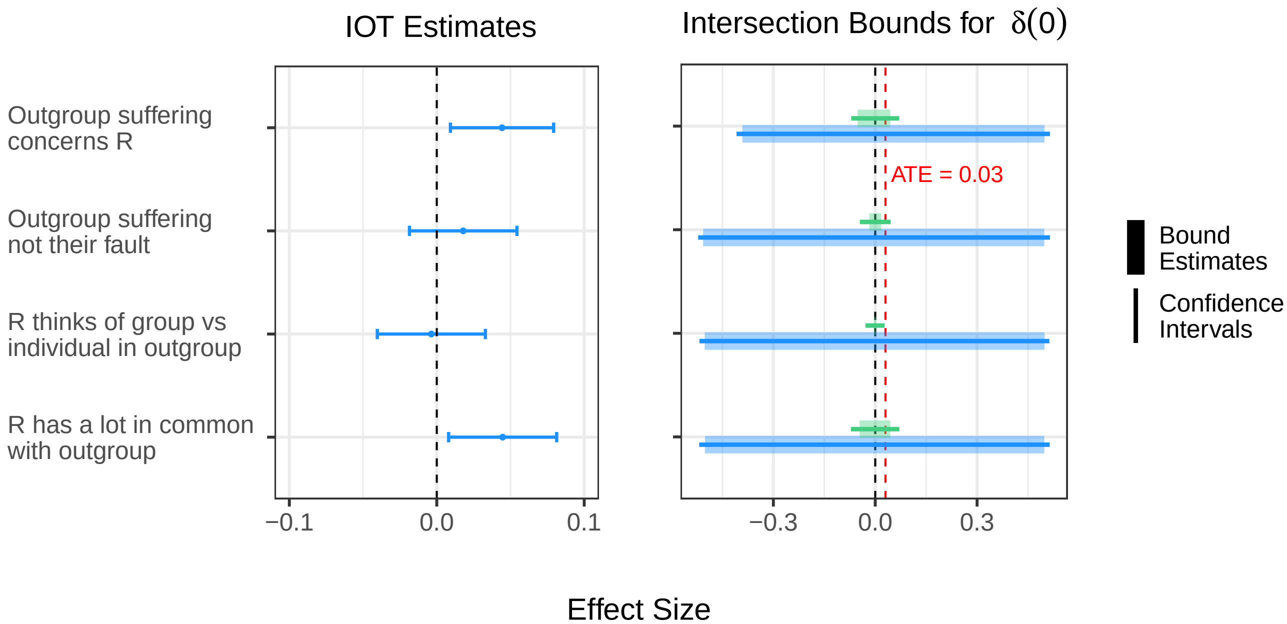

Figure 2: Establishing causal mechanisms for the effect of perspective-getting on prejudice reduction (Kalla and Broockman, 2023). The figure displays IOT point estimates (left panel) and sharp bounds estimates (thick, transparent lines) for the indirect effect of perspective-getting with an essay assignment on prejudice reduction via each of the putative mediators at reference level $A = 0$ ($\delta(0)$), as well as the 95% confidence intervals for these bounds (thin lines). No-assumption bounds are shown in blue, while MMR bounds derived under Assumption 3) are shown in green. The dotted black line marks zero, while the dotted red line marks the ATE (0.03).

For both the no-assumption bounds (blue-colored bounds) and the MMR bounds (green-colored bounds), we present the sharp bounds (thick, transparent lines) in addition to their 95% confidence intervals (thin lines). The estimated no-assumption bounds are extremely wide and approximately center around 0, providing little to no information about the existence of an indirect effect for each of the mediators considered, nor their direction or magnitude. In particular, the no-assumption bounds estimated for the key mediator cited by Kalla and Broockman (2023) – reactive empathy – is $[-0.39, 0.50]$, an interval that contains a large range of positive and negative effect sizes, contains zero, and includes a set of indirect effect sizes on the order of 10 times larger in absolute value than the estimate recovered by the IOT approach. In short, the no-assumption bounds do not allow us to conclude that any of the four mediators considered mediate the effect of perspective-getting on prejudice reduction, even those mediators that "pass" an IOT (i.e., those mediators that are themselves affected by the treatment), such as reactive empathy and attributional thinking.

Does assuming mediator monotonicity – that the treatment affects the mediator in one direction for *all* respondents – enable us to recover the informativeness of the IOT approach? The green bands in Figure 2 panel B address this question. Two points are of particular note. First, the bounds shrink substantially compared with those computed without restriction on the distribution of potential outcomes. For example, the bounds estimate for reactive empathy shrinks from $[-0.39, 0.50]$ under no assumption to $[-0.05, 0.043]$ under MMR, a considerably smaller interval that ranges from approximately $-ATM$ to $+ATM$. Second, and despite this, for the two putative mediators with non-zero IOT estimates (concern about outgroup suffering and feeling commonality with the outgroup), the bounds are uninformative about either the existence or the sign of the relevant indirect effect. The bounds all contain zero and a large range of positive and negative values, including the overall ATE (red dotted line). In other words, the observed data are consistent with the indirect effect via these mediators being zero, as well as with the indirect effect via these mediators being equal to the ATE (and thus, consistent with these mediators serving as the *sole* mechanism through which the ATE operates). Only for the group vs. individual mediator – and to a lesser extent, the "outgroup suffering is not their fault" mediator – is the MMR bounds estimate extremely narrow: in this case, it almost collapses on zero, a result which mirrors the IOT finding of an extremely small effect of treatment on those mediators. This result highlights the 'approximately necessary" condition of IOTs under MMR, in that small ATMs imply small indirect effects.

Is the MMR a reasonable assumption in this setting? For the group vs. individual mediator proposed by Kalla and Broockman (2023), this means assuming that no individual for whom hearing about the feelings of prejudice of trans people or immigrants would make them more likely to view them as individuals rather than as a group. Given the amorphous nature of this kind of evaluation, it seems difficult to maintain this strong of an assumption.

5.2 Transitional Justice Museums and Support for Democracy

We now turn to the field experiment of Balcells, Palanza and Voytas (2022) on transitional justice museums. Around 500 university students in Santiago, Chile, were randomly assigned to either a treatment or control condition. Students in the control group were issued an end-line survey with questions about their political attitudes; students in the treatment group participated in a one-hour tour of the Museum of Memory and Human Rights and then immediately after completed a survey analogous to that administered to the control group. The authors find that museum visitation significantly increased students' support for democratic institutions and transitional justice policies and decreased support for repressive institutions.

The authors posit that museums' narratives of past events may have elicited emotional reactions among their visitors. To test this hypothesis, they measured treated and untreated students' responses to the Positive and Negative Affect Schedule, which asks respondents to consider a range of positive and negative emotions and to indicate which they feel in the present moment (Watson, Clark and Tellegen, 1988). We focus on the negative emotion questions as the authors find treatment effects on these emotions to be substantively larger and more consistent than the positive emotion questions. Similar to the previous example, we dichotomize all mediators as well as the outcome variable at their medians. Moreover, given the randomized nature of treatment in this study, we estimate the IOT effects using bivariate regression between treatment and mediator, similar to the previous application. The left panel of Figure 3 reports the IOT estimates, which replicate the patterns reported by Balcells, Palanza and Voytas (2022): the visit to the transitional justice museum has a positive and statistically significant effect on the overall negative emotions experienced by respondents (0.46, 95% confidence interval of $[0.35, 0.57]$). In particular, the authors find positive and significant treatment effects on feeling tense, scared, guilty, hostile, fearful, afraid, and embarrassed. Balcells, Palanza and Voytas (2022) argue that these findings indicate that the emotional experience of a transitional justice museum may act as a key mechanism by which the museum visit alters political attitudes.

While the IOTs could indicate that the strong and significant emotional reactions elicited by a museum visit could serve as an important mediating pathway for the overall museum effect, the right

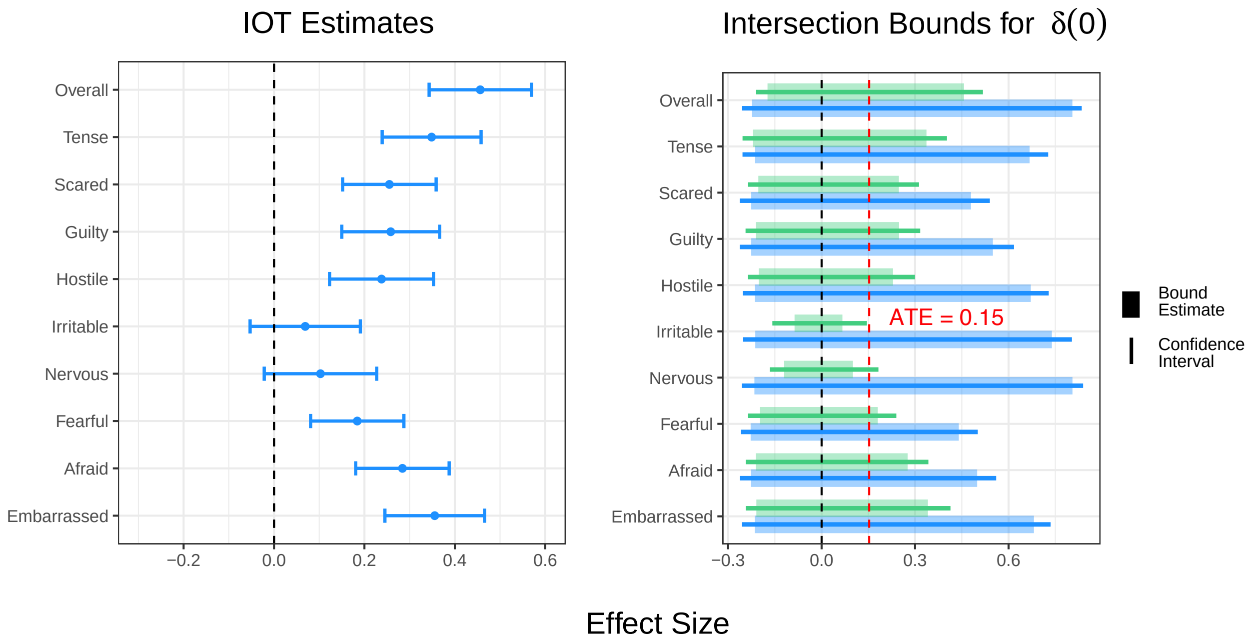

Figure 3: Establishing causal mechanisms for the effect of a transitional justice museum visit on support for democratic institutions (Balcells, Palanza and Voytas, 2022). The figure displays IOT point estimates (left panel) and sharp bounds estimates (thick, transparent lines) for the indirect effect of the museum intervention via each of the putative mediators at reference level $A = 0$ ($\delta(0)$), as well as the 95% confidence intervals for these bounds (thin lines). No-assumption bounds are shown in blue, while MMR bounds derived under Assumption 3) are shown in green. The dotted black line indicates zero, the dotted red line indicates the ATE (0.15) of the museum visit on support for democracy.

panel of fig. 3 cautions against such an interpretation. Paralleling the canvassing example, we make three remarks about the bounds on $\delta(0)$. First, all no-assumption bounds are extremely wide, and all contain zero. For example, the bounds for the overall negative emotion outcome is [-0.22, 0.80]. Second, the MMR bounds are significantly shorter than their no-assumption counterparts, although they still contain zero. Therefore, while the MMR assumption seems relatively plausible under this empirical setting—that compared to the control condition, visiting a museum recounting the country's past political violence only increases negative emotions such as fear and guilt—the bounds derived under this assumption remain uninformative about mediation via each of the variables considered. Finally, smaller IOT coefficients correspond to shorter bounds under the MMR assumption. For instance, the effect of the museum visit on "feeling irritable" has a point estimate of 0.07, the smallest among all estimates. Correspondingly, the bound for "feeling irritable" is also the shortest, ranging from $[-0.09, 0.07]$, which is approximately $-ATM$ to $+ATM$. Even for these mediators, however, the confidence intervals are close to the overall ATE, meaning we cannot rule out the possibility that they account for the entire effect of the museum intervention.

6 What Is the Point of Mechanism Tests?

As these examples show, the question of mechanisms has become an important part of empirical political science. Indeed, according to our literature review, over a third of recent empirical papers in top political science journals deploy some form of mechanism testing. Why is there an emphasis on this project if many believe the assumptions necessary to sustain such tests are implausible?

In the social sciences, the idealized version of empirical research happens by a process where researchers (a) develop a relatively broad causal theory (or explanation), (b) use this theory to derive specific hypotheses about specific causal effects, and (c) use convincing causal identification and suitable data and estimators to test these hypotheses. One hopes the connection between a hypothesis and the estimated causal effect will be strong, but one hypothesis (and the resulting causal effect) may be consistent with several broader causal theories. Researchers hope to use mechanism tests to bridge this gap by showing the path by which the treatment affects the outcome. A causal mechanism helps generalize a specific causal effect into a more general form, allowing us to hopefully better understand all situations in which this mechanism is activated. Unfortunately, there are many instances where all approaches to establishing indirect effects, including IOTs, rely on shaky assumptions that are difficult to justify. What can be done in that case?

We encourage researchers to pursue the time-tested approach of "implication analysis" (Lieberson and Horwich, 2008) or "inference to the best explanation" (Spirling and Stewart, 2024), whereby researchers generate many different hypotheses that differentiate between the proposed theory and alternatives. The benefit of this approach is that these secondary hypotheses can be completely unrelated to the direct or indirect effects of the main causal effect of interest. This can free researchers from focusing on research designs with assumptions ill-suited to a particular setting. Of course, as we have articulated here, we must be clear about the assumptions needed to draw conclusions from these tests. In many ways, the main goal of Kalla and Broockman (2023) takes this exact approach. Earlier work showed that interpersonal conversations can lead to reductions in exclusionary attitudes, but it was not clear *why* these conversations worked. The experimental design of Kalla and Broockman (2023) is tailored to test different types of narrative strategies to help adjudicate between different theoretical mechanisms of different narrative strategies. Thus, their main analysis helps clarify causal mechanisms without resorting to IOTs, effect heterogeneity, or causal mediation.

7 Conclusion

This paper shows how a popular way of assessing causal mechanisms—the intermediate outcome test—relies on strong and often undisclosed assumptions. In these tests, authors present the average effects of treatment on the mediator and often argue that these effects show plausible causal pathways of the overall effect of the treatment on the outcome. However, our analysis shows that IOTs can neither establish nor rule out indirect effects through a mediator under randomization of treatment alone. We showed that IOTs *can* rule out indirect effects if the effect of treatment on the mediator can only move units in one direction, which is a very strong monotonicity assumption. Furthermore, IOTs cannot establish an indirect effect even if we assume a positive average effect of the mediator on the outcome. We derive sharp nonparametric bounds for the indirect effect under these various assumptions and estimate these bounds in two empirical applications. Finally, we also discussed how causal mechanisms might be assessed in situations where the assumptions needed to make inferences about indirect effects are questionable.

We believe there is much future research to be completed in the field of causal mechanisms. In particular, future research should attempt to unify different accounts of "mechanisms" in a single broad framework. Currently, mechanisms are sometimes used interchangeably with "indirect effects," but other times, scholars use mechanisms to refer to a theoretical construct that can explain an effect without reference to any mediation at all. In epidemiology, there is a tradition of defining causal mechanisms in the "sufficient cause framework," where authors seek to find sets of variable levels that are sufficient to produce a particular outcome (VanderWeele, 2009). Connecting the ideas of mediation and mechanism to the theoretical underpinnings of the social sciences would help scholars develop better empirical tests of their theories.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3):512–529.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2018. "Analyzing Causal Mechanisms in Survey Experiments." *Political Analysis* 26(4):357–378.

Adida, Claire L, Adeline Lo and Melina R Platas. 2018. "Perspective Taking Can Promote Short-term Inclusionary Behavior Toward Syrian Refugees." *Proceedings of the National Academy of Sciences* 115(38):9521–9526.

Audette, Nicole, Jeremy Horowitz and Kristin Michelitch. 2020. Personal Narratives Reduce Negative Attitudes Toward Refugees and Immigrant Outgroups: Evidence from Kenya. Working Paper 1-2020 Vanderbilt University Center for the Study of Democratic Institutions.
URL: *https://www.vanderbilt.edu/csdi/includes/WP_1_2020_FINAL.pdf*

Balcells, Laia, Valeria Palanza and Elsa Voytas. 2022. "Do Transitional Justice Museums Persuade Visitors? Evidence from a Field Experiment." *The Journal of Politics* 84(1):496–510.

Balke, Alexander and Judea Pearl. 1997. "Bounds on Treatment Effects from Studies with Imperfect Compliance." *Journal of the American Statistical Association* 92:1171–1176.

Blackwell, Matthew and Anton Strezhnev. 2022. "Telescope Matching for Reducing Model Dependence in the Estimation of the Effects of Time-varying Treatments: An Application to Negative Advertising." *Journal of the Royal Statistical Society Series A: Statistics in Society* 185(1):377–399.

Broockman, David and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-door Canvassing." *Science* 352(6282):220–224.

Bruneau, Emile G and Rebecca Saxe. 2012. "The Power of Being Heard: The Benefits of 'Perspective-giving' in the Context of Intergroup Conflict." *Journal of Experimental Social Psychology* 48(4):855–866.

Bullock, John G. and Donald P. Green. 2021. "The Failings of Conventional Mediation Analysis and a Design-Based Alternative." *Advances in Methods and Practices in Psychological Science* 4(4):1–18.

Callis, Anna, Thad Dunning and Guadalupe Tuñón. 2024. Causal Inference and Knowledge Accumulation in Historical Political Economy. In *The Oxford Handbook of Historical Political Economy*. Oxford University Press pp. 55–74.

Chernozhukov, Victor, Sokbae Lee and Adam M. Rosen. 2013. "Intersection Bounds: Estimation and Inference." *Econometrica* 81(2):667–737.

Daniel, Rhian M, Bianca L De Stavola, Simon N Cousens and Stijn Vansteelandt. 2015. "Causal Mediation Analysis with Multiple Mediators." *Biometrics* 71(1):1–14.

De Brito, Alexandra Barahona, Carmen González Enríquez and Paloma Aguilar. 2001. *The Politics of Memory: Transitional Justice in Democratizing Societies*. Oxford University Press.

Fu, Jiawei and Tara Slough. 2024. "Heterogeneous Treatment Effects and Causal Mechanisms." arXiv preprint arXiv:2404.01566. Version: 15 Jun 2024.
URL: *https://arxiv.org/abs/2404.01566*

Galinsky, Adam D and Gordon B Moskowitz. 2000. "Perspective-taking: Decreasing Stereotype Expression, Stereotype Accessibility, and In-group Favoritism." *Journal of Personality and Social Psychology* 78(4):708.

Glynn, Adam N. 2012. "The Product and Difference Fallacies for Indirect Effects." *American Journal of Political Science* 56(1):257–269.

Goetgeluk, Sylvie, Stijn Vansteelandt and Els Goetghebeur. 2008. "Estimation of Controlled Direct Effects." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70(5):1049–1066.

Green, Donald P., Shang E. Ha and John G. Bullock. 2010. "Enough Already about "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose." *The ANNALS of the American Academy of Political and Social Science* 628(1):200–208.

Hirano, Keisuke and Jack R. Porter. 2012. "Impossibility Results for Nondifferentiable Functionals." *Econometrica* 80(4):1769–1790.

Horne, Cynthia M. 2014. "The Impact of Lustration on Democratization in Postcommunist Countries." *International Journal of Transitional Justice* 8(3):496–521.

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

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

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

Imai, Kosuke and Teppei Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21(2):141–171.

Kalla, Joshua L and David E Broockman. 2023. "Which Narrative Strategies Durably Reduce Prejudice? Evidence from Field and Survey Experiments Supporting the Efficacy of Perspective-getting." *American Journal of Political Science* 67(1):185–204.

Levis, Alexander W., Matteo Bonvini, Zhenghao Zeng, Luke Keele and Edward H. Kennedy. 2023. "Covariate-assisted Bounds on Causal Effects with Instrumental Variables." arXiv preprint

arXiv:2301.12106. Version: 29 Sep 2023.
URL: *https://arxiv.org/abs/2301.12106*

Lieberson, Stanley and Joel Horwich. 2008. "Implication Analysis: A Pragmatic Proposal for Linking Theory and Data in the Social Sciences." *Sociological Methodology* 38(1):1–50.

Lowe, Matt. 2021. "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration." *American Economic Review* 111(6):1807–1844.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Harvard University Press.

Nalepa, Monika. 2010. *Skeletons in the Closet: Transitional Justice in Post-communist Europe*. Cambridge University Press.

Pearl, Judea. 2001. Direct and Indirect Effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, ed. J.S. Breese and D. Koller. Morgan Kaufmann Publishers San Francisco, CA: pp. 411–420.

Sachs, Michael C, Gustav Jonzon, Arvid Sjölander and Erin E Gabriel. 2023. "A General Method for Deriving Tight Symbolic Bounds on Causal Effects." *Journal of Computational and Graphical Statistics* 32(2):567–576.

Simonovits, Gábor, Gabor Kezdi and Peter Kardos. 2018. "Seeing the World Through the Other's Eye: An Online Intervention Reducing Ethnic Prejudice." *American Political Science Review* 112(1):186–193.

Spirling, Arthur and Brandon M. Stewart. 2024. "What Good Is A Regression? *Inference to the Best Explanation* and the Practice of Political Science Research." *Journal of Politics* p. Forthcoming.

VanderWeele, Tyler J. 2009. "Mediation and Mechanism." *European Journal of Epidemiology* 24(5):217–224.

VanderWeele, Tyler J, Stijn Vansteelandt and James M Robins. 2014. "Effect Decomposition in the Presence of an Exposure-induced Mediator-outcome Confounder." *Epidemiology* 25(2):300–306.

Watson, David, Lee Anna Clark and Auke Tellegen. 1988. "Development and Validation of Brief Measures of Positive and Negative affect: the PANAS scales." *Journal of Personality and Social Psychology* 54(6):1063.

Williamson, Scott, Claire L Adida, Adeline Lo, Melina R Platas, Lauren Prather and Seth H Werfel. 2021. "Family Matters: How Immigrant Histories Can Promote Inclusion." *American Political Science Review* 115(2):686–693.

Zhou, Xiang and Geoffrey T Wodtke. 2019. "A Regression-with-residuals Method for Estimating Controlled Direct Effects." *Political Analysis* 27(3):360–369.

Zhou, Xiang and Teppei Yamamoto. 2023. "Tracing Causal Paths from Experimental and Observational Data." *The Journal of Politics* 85(1):250–265.

Supplemental Materials (to appear online)

A Overview of the linear programming approach

The bounds we have derived in the main text come from an optimization problem that can be solved via a linear programming approach. This approach largely follows the discussion of Imai, Tingley and Yamamoto (2013) with different maintained assumptions. For each indirect effect, one of the values in the causal contrast is identified. For instance, with $\overline{\delta}(1)$, we only need to consider $\mathbb{E}[Y_i(1, M_i(0))]$ since $\mathbb{E}[Y_i(1, M_i(1))] = \mathbb{E}[Y_i(1)] = \mathbb{E}[Y_i \mid A_i = 1]$ is identified without further assumptions. The goal of the partial identification, then, is to find the maximum and minimum value of $\mathbb{E}[Y_i(1, M_i(0))]$ that is consistent with the data and the maintained assumptions. To do this, we introduce an augmented set of principal strata,

$$\psi_{y_1 y_0 m_1 m_0} = \mathbb{E}[Y_i(1,1) = y_1, Y_i(1,0) = y_0, M_i(1) = m_1, M_i(0) = m_0].$$

Note that we only need to consider the potential outcomes under treatment $Y_i(1, m)$ here since they are the only quantities that appear in the ANIE.

We can write the subject of optimization, the cross-world counterfactual, in terms of the principal strata as

$$\mathbb{E}[Y_i(1, M_i(0))] = \sum_{y=0}^{1} \sum_{m=0}^{1} \psi_{1ym1} + \psi_{y1m0} \tag{10}$$

There are several constraints that must hold on the principal strata dictated by the axioms of probability or assumptions about how the observed data and potential outcomes relate. First, we have the logical constraint

$$\sum_{y_1=0}^{1} \sum_{y_0=0}^{1} \sum_{m1=0}^{1} \sum_{m0=0}^{1} \psi_{y_1 y_0 m_1 m_0} = 1,$$

and the relationship between the observed probabilities and the principal strata,

$$p_{ym\cdot 1} = \begin{cases} \sum_{y_0=0}^{1} \sum_{m_0=0}^{1} \psi_{y y_0 m m_0} & \text{if } m = 1 \\ \sum_{y_1=0}^{1} \sum_{m_0=0}^{1} \psi_{y_1 y m m_0} & \text{if } m = 0 \end{cases}.$$

These two restrictions (along with a nonnegativity constraint on all ψ values) are the only restrictions from just randomization. Assumption 3 (MMR) implies that any mechanism defier strata has zero probability of occurring,

$$\psi_{y_1 y_0 01} = 0 \qquad \forall y_1 \in \{0, 1\}, y_0 \in \{0, 1\}.$$

Finally, for Proposition 2, we can add the restriction that the effect of $\mathbb{E}[Y_i(1, 1) - Y_i(1, 0)] \geq 0$ with

$$\sum_{m_1=0}^{1} \sum_{m_0=0}^{1} \psi_{10 m_1 m_0} - \sum_{m_1=0}^{1} \sum_{m_0=0}^{1} \psi_{01 m_1 m_0} \geq 0.$$

Using standard linear programming techniques, we can obtain bounds by maximizing or minimizing (10) with respect to these restrictions. We can then combine these with the already identified expressions for $\mathbb{E}[Y_i(1, M_i(1))]$ to obtain sharp bounds for $\overline{\delta}(1)$. A similar approach can obtain bounds for $\overline{\delta}(0)$.