Extract Mechanisms from Heterogeneous Effects: A New Identification Strategy for Mediation Analysis [*]

Jiawei Fu[†]

November 11, 2023

Abstract

Understanding causal mechanisms is essential for explaining and generalizing empirical phenomena. Causal mediation analysis offers statistical techniques to quantify mediation effects. However, existing methods typically require strong identification assumptions or sophisticated research designs. We develop a new identification strategy that simplifies these assumptions, enabling the simultaneous estimation of causal and mediation effects. The strategy is based on a novel decomposition of total treatment effects, which transforms the challenging mediation problem into a simple linear regression problem. The new method establishes a new link between causal mediation and causal moderation. We discuss several research designs and estimators to increase the usability of our identification strategy for a variety of empirical studies. We demonstrate the application of our method by estimating the causal mediation effect in experiments concerning common pool resource governance and voting information. Additionally, we have created statistical software to facilitate the implementation of our method.

Keywords: Mediation Analysis, Identification, Heterogeneous Treatment Effects, Mechanism, Moderation

[*] I thank Neal Beck, Kouske Imai, Connor Jerzak, Cyrus Samii, Yuki Shiraito, Tara Slough, Teppei Yamamoto; seminar audiences at New York University; and participants at PolMeth XL and APSA 2023 3for helpful feedback.

[†] Ph.D. Candidate, New York University jf3739@nyu.edu

1 Introduction

"How and why does the treatment affect the outcome?" Many answers from different angles exist. Causal mediation analysis traces the underlying mechanical process and quantifies the effect that is mediated through some intermediate variables between the treatment and the outcome (Imai et al. 2011; VanderWeele 2015). However, practical application of mediation analysis is frequently hindered by the implausibility of onerous assumptions. This study proposes a novel and straightforward identification strategy to facilitate the causal mediation analysis. In general, our strategy can convert a difficult causal problem into a simple regression problem, which allow researchers estimate causal mediation effect and treatment effect simultaneously.

Figure 1 illustrates some basic concepts of mediation analysis. On the left panel, the total treatment effect is decomposed into direct effect and indirect effect. Because of the existence of a mediator in the indirect effect, mediation analysis requires stronger assumptions than what is required for the identification of total treatment. In the language of causal inference, notably, the identification of treatment effects only requires to address the confounding between the treatment and the outcome (U_3 in the right panel), while mediation analysis requires the ignorability of the mediator (U_2 and U_4). In a typical empirical study, satisfying and justifying these two assumptions is challenging. Ideally, a causal mediation identification strategy would simultaneously estimate the total treatment and mediation effects, or at least not impose additional identification burdens. Our new method can achieve this goal.

Traditionally, mediation analysis has been approached through two methods. Over the decades, path analysis and structural equation modeling (SEM) have been the most widely used tools for mediation analysis in social science. Inspired by Baron and Kenny (1986), scholars typically employ linear regression models and utilize the "product" or

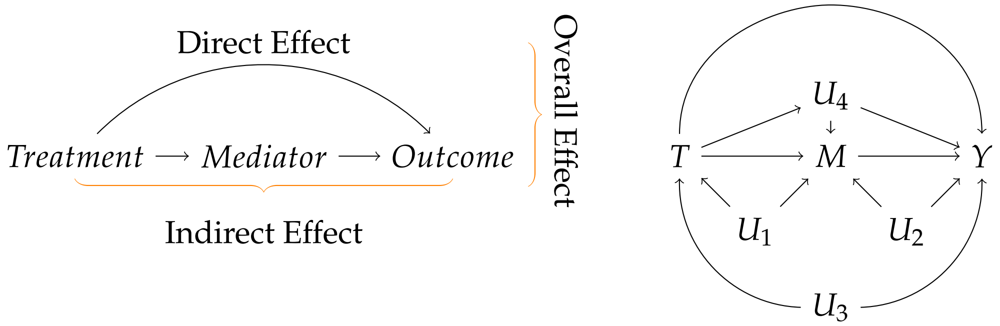

Figure 1: The left panel illustrates the basic decomposition of the overall treatment effect. The right panel is one example of the directed acyclic graph. T is the treatment variable, M is the mediator, Y is the outcome variable, and $U_j, j = 1, 2, 3, 4$ are confounders.

"difference" estimators. Through structural equations, this approach explicitly illustrates the mechanical process by which treatment affects the mediator and subsequently how the mediator affects the outcome, although it involves several parametric assumptions. To identify coefficients, similar to ordinary least squares (OLS) estimator, strong 'no omitted variables' assumptions are required.

Conversely, the counterfactual approach emphasizes the causal interpretation of mediation employing a more flexible potential outcome framework. By decomposing the overall treatment effect into natural direct and indirect effects, this approach deepens our understanding of the causal mechanism and clarifies unconfoundness assumptions. However, as discussed in section 2, the proposed decomposition understates the mechanical process from treatment to mediator, thus restricting us from developing new identification strategies. Parallel to causal inference, non-parametric identification of mediation effects requires sequential ignorability (Imai et al. 2010a). Besides the treatment being ignorable, the observed mediator must also be ignorable, given the actual treatment status. The second ignorability requirement is the primary obstacle to more widespread application.

Our novel identification strategy incorporates the strength of two approaches and does not require strong exogeneity assumptions. Under the counterfactual approach, we first decompose the total treatment effect by highlighting the treatment effect on the mediator. As we show in section 4, the new decomposition has the form of simple linear regression, where the 'dependent variable' is the average treatment effect on the outcome, the 'independent variable' is the average total treatment effect on the mediator, and the 'slope' is the natural indirect effect. Therefore, if the slope is identified, we are able to identify the causal mediation effect. As in OLS, the critical identification assumption is that the average treatment effect on the mediator is not correlated with the error term. In our decomposition, the variance within the error term comes from other mechanisms. Under this assumption, even the simplest OLS estimator can recover the causal mediation effect. It is worth mentioning that this is not an entirely new identification assumption. As far as I know, it is explored in the context of invalid instrumental variables (Kolesár et al. 2015) and Mendelian randomization (Bowden et al. 2015). Nevertheless, this is the first instance of its use in the context of mediation analysis.

To implement the strategy, the remaining question pertains to obtaining multiple observed average treatment effects on the outcome and the mediator. It must be obtained from the treatment heterogeneity. In section 5, we propose three possible research designs. In the first design, we assume the existence of subgroups in the population that have heterogeneous treatment effects. We suggest researchers use pre-treatment covariates to identify those subgroups, probably through theory or data-driven methods like causal tree or forest (Wager and Athey 2018). Next, for each subgroup, we obtain the required data (average treatment effects on the outcome and on the mediator), which we refer to as *Heterogeneous Subgroup Design*. The second design explores multiple types of treatment. For example, how does contact affect turnout? In the Get-Out-The-Vote (GOTV) experiments, researchers consider several contact treatments, including phone calls, email, door-to-door canvass, and many more. Each treatment can be regarded as sub-types of the meta-treatment (contact) and generate distinct average treatment effects, which we refer to as *Multiple Treatment Meta Design*. The third design incorporates both of the previous two designs.

Because average treatment effects are estimated from the design, we need to modify the simple OLS estimator to account for the measurement error. We discuss several estimators. The first class is the aggregate-level estimator that only uses average treatment effects rather than individual data. Therefore, the estimator can allow us to combine results from multiple studies, similar to meta-analysis. The second class is the individual-level estimator, which can provide more precise results by assuming some parametric structure. In section 6, we use Monte Carlo simulation to examine the performance of those estimators, especially in the context of small sample sizes. The results are promising, showing effectiveness even with minimal data points, such as five observations. We then apply our methodology to estimate causal mediation effects in two empirical studies using actual data. The first study is about 'Governance on Resources'. We demonstrate the estimation of mediation effects using just six estimates derived from six different sites. The second examines the impact of information on voting behavior, where we calculate the mediation effect using individual-level data. Overall, provided that treatment effects are identifiable, researchers can ascertain causal mediation effects with relative ease under reasonable assumptions through our proposed approach.

Our identification strategy turns out to be extremely simple; it does not require any advanced techniques beyond simple linear regression. One reason is that it exploits the casual heterogeneity, which is commonly ignored in causal mediation analysis literature. Although causal mediation and causal moderation are active research areas that have frequently been discussed together (Baron and Kenny 1986), they seldom are united in the same discussion. Causal heterogeneity should provide abundant information on causal mechanisms. In practice, researchers frequently use heterogeneous treatment effects (HTE) to detect mechanism activation. Fu and Slough (2023) develop a theoretical framework to link HTE and causal mediation; they clarify the underlying identification assumptions and discuss the potential limitations. In contrast, our new method attempts to use HTE to quantify the mediation effect directly. We believe that extracting mechanical information from causal moderation can provide many more interesting perspectives on mediation analysis.

2 Causation and Mediation

The discussion on causation and mediation generally takes two different approaches: the counterfactual and structural approaches. Our new decomposition incorporates the relative advantages in both approaches. Therefore, in this section, we briefly introduce and compare them, if necessary, with the help of directed acyclic graphs (DAG).[1] The causal decomposition under either approach has its own advantages and disadvantages. Our critical comparisons indicate a potential synthetic approach for the new identification strategy.

2.1 Counterfactual Approach

In the counterfactual approach of causal inference, causation and mediation are interpreted and decomposed with potential outcomes (Holland 1986). Let T be the binary treatment and Y be the outcome variable. Suppose there exist J independent mediators M_j. The overall treatment effect of T on Y for individual i, denoted by τ^i, is represented

[1] Although there exist some debates about the potential outcome approach and graphic approach (see Imbens (2020), and Perl's reply: http://causality.cs.ucla.edu/blog/index.php/2020/01/29/on-imbens-comparison-of-two-approaches-to-empirical-economics/), we still find both approaches have their own particular merits.

by

$$\tau^i = Y^i(1, M_1^i(1), ..., M_j^i(1)) - Y^i(0, M_1^i(0), ..., M_j^i(0)).$$

For the total effects, all mediators should consider potential outcomes under the treatment status. Following Pearl (2001) and Robins and Greenland (1992), total treatment effects can be decomposed into natural direct and indirect effects. The natural direct effect for individual i (δ^i) compares the outcome under treatment and control groups, but mediators are set to be potential outcomes under a specific treatment assignment $t' = 0$ or 1 [2],

$$\delta^i(t') = Y^i(t', M_1^i(t'), ..., M_J^i(t')) - Y^i(t, M_1^i(t'), ..., M_J^i(t')).$$

The terminology "natural" is in contrast to "controlled." For controlled direct effect, we fix the mediator at a certain value m_j, rather than their potential outcomes under a given treatment assignment. Therefore, the controlled direct effect can be defined as $Y^i(1, m_1, ..., m_J) - Y^i(0, m_1, ..., m_J)$ (see Acharya et al. 2016).

The natural indirect effect through mediator j for individual i (η_j^i) reflects the effect on the outcome by changing the mediator j. Because we allow for multiple mechanisms, we depart from the convention by using $j-$ and $j+$ to denote index $h \in J$ such that $h < j$ and $h > j$, respectively. For $t, t' = 0$ or 1, the natural indirect effect is defined as

$$\begin{aligned}\eta_j^i(t', t) = Y^i(t, M_{j-}(t), M_j(t'), M_{j+}(t')) - \\ Y^i(t, M_{j-}(t), M_j(t), M_{j+}(t'))\end{aligned}$$

As is standard, we can then re-write the total causal effect the sum of natural direct and indirect effects [3]:

[2] The definition in the main text is simplified; for a complete discussion, see supplementary materials.

[3] The intuition for our notation is as follows. Define $\eta_0(t', t) = \delta(t')$; The first term of $\eta_j(t', t)$ and the second term of $\eta_{j-1}(t', t)$ cancel out. See SI A.

$$\tau^i = \delta^i(t') + \sum_{j=1}^{J} \eta_j^i(t', t)$$

Due to the fundamental problem of causal inference, we are confined to the estimation of aggregate effects, typically averages. We therefore define τ as the average overall treatment effect: $\tau := \mathbb{E}\tau^i$, δ as the average direct effect: $\mathbb{E}\delta^i$, and η_j as the average indirect effect for mechanism j: $\eta_j := \mathbb{E}\eta_j^i$.

For simplicity, we only consider a single mediator so that the subscript j is omitted. All results of the study can be naturally extended to multiple mechanisms, which are discussed in supplementary materials.

It is evident that multiple versions of natural direct and indirect effects exist. The main concern is the interaction effect between the treatment and the mediator. For natural direct effect $\delta^i(t) = Y^i(1, M^i(t)) - Y^i(0, M^i(t))$ and natural indirect effect $\eta^i(t) = Y^i(t, M^i(1)) - Y^i(t, M^i(0))$, the value of $\delta^i(t)$ and $\eta^i(t)$ may depend on t, if there exists an interaction effect. In the binary treatment case ($t = 1$ and $t' = 0$), they have specific names (Robins and Greenland 1992):

(1) The "pure" effect implies that no interaction effect is picked up. We call $\delta^i(0) = Y^i(1, M^i(0)) - Y^i(0, M^i(0))$ the *pure direct effect*, where the mediator is set to the value it would have been without treatment. Additionally, $\eta^i(0) = Y^i(0, M^i(1)) - Y^i(0, M^i(0))$ is the *pure indirect effect* where treatment is set to absent.

(2) The "total" effect captures the interaction effect. Therefore the *total direct effect* is defined as $\delta^i(1) = Y^i(1, M^i(1)) - Y^i(0, M^i(1))$, where the mediator takes the potential value if the treatment is on. Similarly, the *total indirect effect* $\eta^i(1) = Y^i(1, M^i(1)) - Y^i(1, M^i(0))$ set treatment to present.

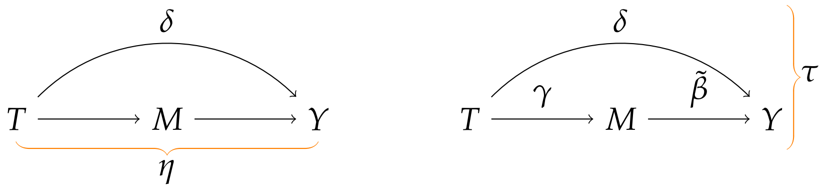

Figure 2: Decompositions with Counterfactual and Structural Approaches

Together, we obtain two different decompositions:

$$\begin{aligned}\tau &= \delta(0) + \eta(1) \\ \tau &= \delta(1) + \eta(0)\end{aligned} \tag{1}$$

Assuming no interaction effect exists between the treatment and mediator, the natural and total direct (indirect) effects should be the same because the effect does not depend on the mediator (treatment). Under the assumption, $\delta = \delta(0) = \delta(1)$ and $\eta = \eta(0) = \eta(1)$. Then the decomposition is unique: $\tau = \delta + \eta$. The left DAG in Figure 2 illustrates this basic structure.

If people do not assume "no interaction effect", we can further decompose the total direct effect or total indirect effect to emphasize the interaction effect. VanderWeele (2013) proposes further decomposing the total direct (or indirect) effect. For example, suppose the mediator is binary, then the total indirect effect can be decomposed into pure indirect effect and the interaction term: $\eta(1) = \eta(0) + [(Y(1, M(1)) - Y(1, M(0)) - Y(0, M(1)) + Y(0, M(0))](M(1) - M(0))$.

Notably, other decomposition methods within the counterfactual approach exist. For example, the study by Gallop et al. (2009) examines mediation analysis under principal strata. However, as emphasized by VanderWeele (2011), decomposition based on principal strata does not equate to the natural direct effect and natural indirect effect. Moreover, membership in principal strata themselves generally are unidentified.

2.2 Structural Approach

Historically, path analysis and structural equation modeling (SEM) is the predominantly used framework for conducting mediation analysis (see Hong 2015; MacKinnon 2012). As a special case, Baron and Kenny (1986) develop the linear additive model using a single treatment, mediator, and outcome variable. It comprises two main equations:

$$Y = \alpha_1 + \delta T + \tilde{\beta} M + \varepsilon_1 \tag{2}$$

$$M = \alpha_2 + \gamma T + \varepsilon_2 \tag{3}$$

Replacing M (equation (3)) in (2), we obtain the reduced form as follows:

$$Y = (\alpha_1 + \alpha_2 \tilde{\beta}) + (\delta + \tilde{\beta}\gamma)T + (\varepsilon_1 + \tilde{\beta}\varepsilon_2) \tag{4}$$

$$:= \alpha_3 + \tau T + \varepsilon_3 \tag{5}$$

Traditionally, parameter before the treatment T, i.e., $\tau = \delta + \tilde{\beta}\gamma$ in (5) is interpreted as the total treatment effect; δ is interpreted as the direct effect; and $\tilde{\beta}\gamma$ is interpreted as the indirect effect. The right part of figure 2 illustrates this DGP.

Unlike the counterfactual approach, structural models implicitly incorporate parametric assumptions, such as linearity and constant effects. However, this representation highlights the mechanical process through which the treatment is mediated by intermediate variables. It explicitly indicates that the indirect effect has two components, which is not emphasized in the counterfactual approach. The first component, γ, is the effect from the treatment to the mediator, and the second component, $\tilde{\beta}$, is the effect from the mediator to the outcome. In the potential outcome representation, however, $\eta(t) = Y(t, M(1)) - Y(t, M(0))$ conflates two effects, which obscures the mechanical pathways of how the treatment is mediated by intermediate variables. As demonstrated below, the re-introduction of γ into the counterfactual decomposition, generates new insights about the mediation analysis.

3 Identification Assumptions

Before demonstrating our new identification strategies, it is crucial to review the current state of the art. We start with non-parametric and model-based identification. Subsequently, we provide a brief overview of several identification strategies that aid in ensuring the validity of the fundamental identification assumptions. The comparison of our method to current methods will be provided later.

3.1 Non-parametric Assumptions

To non-parametrically identify the causal mediation effect, we require the sequential ignorability. In contrast to the causal inference literature, multiple versions of sequential ignorability assumption exist.[4]. One of the most concise versions is given by Imai et al. (2010b) Formally, it has two important parts. The first part is similar to the unconfoundedness assumption in causal inference. Essentially, it requires treatment assignment to be ignorable given the observed pretreatment confounders X:

$$\{Y^i(t', m), M^i(t)\} \perp\!\!\!\perp T^i | X^i = x \tag{6}$$

Notably, the treatment value is different for outcome Y and mediator M. Hence, it specifies the full joint distribution of all the potential outcome and mediator variables (Ten Have and Joffe 2012). The second part entails the mediator is ignorable given the

[4] In causal inference, the basic identification assumption is the treatment is (conditionally) independent of the potential outcomes $Y^i(t) \perp\!\!\!\perp T^i | X^i$

observed treatment and pre-treatment confounders:

$$Y^i(t', m) \perp\!\!\!\perp M^i(t) | T^i = t, X^i = x \tag{7}$$

In the assumption (7), the mediator takes the value at the "current" treatment assignment t, but the potential outcome is under treatment assignment t'. The different indices make it hard to satisfy in practice. Generally, as mentioned by Imai et al. (2011), an experiment that randomizes both treatment and mediator does not suffice for this assumption to hold. This assumption is used to replace a similar assumption proposed by Pearl (2001): independence between two potential outcomes: $M_t \perp\!\!\!\perp Y_{t',m} | X = x$. This assumption is also strong; it requires cross-world independence [5], which makes it challenging to interpret and cannot be satisfied by any experimental design because we can not let the same individual simultaneously take and do not take the treatment (Pearl 2014).

Under the sequential ignorability, we may non-parametrically identify natural indirect effects:

$$\eta(t|X = x) = \sum_m \mathbb{E}[Y|t, m, x][\mathbb{P}(m|t, x) - P(m|t', x)]$$

The aforementioned formula is the same as the mediation formula, which is derived by Pearl (2001), under a similar but slightly stronger version of (6): $\mathbb{P}(Y(t, m)|X = x)$ and $\mathbb{P}(M(t')|X = x)$ are identified (see Pearl 2014, also).

One important limitation of both assumptions is that all covariates X must be pre-treatment; generally, the natural indirect effect is not identified even if we have data on the post-treatment confounders (Avin et al. 2005).

Several alternative versions of sequential ignorability relax the "cross-world/indices" property and no post-treatment confounders, but require other assumptions. Hafeman

[5] Basically, it requires the independence between two potential outcomes under two different treatment assignments.

and VanderWeele (2011) remove the "cross-world" in (6) to be: $\{Y^i(t,m), M^i(t)\} \perp\!\!\!\perp T^i | X^i = x$. However, to identify the natural indirect effect, they require the mediator to be binary and impose a no-interaction assumption: $\mathbb{E}[Y(1,m) - Y(0,m) | T = 1, M = 1, X = x] = \mathbb{E}[Y(1,m) - Y(0,m) | T = 1, M = 0, X = x]$.

Robins (2003) proposes the finest fully randomized causally interpreted structured tree graph (FRCISTG) model (in contrast to the non-parametric structural equation model for graph). In this semantics of causal DAG, we can relax the "cross-world/indices" property in (7) and allow post-treatment confounders in X: $Y^i(t,m) \perp\!\!\!\perp M^i(t) | T^i = t, X^i = x$. Again, another no-interaction assumption is required to non-parametrically identify causal mediation effect: $Y^i(1,m) - Y^i(0,m) = Y^i(1,m') - Y^i(0,m') = B(t,t')$ where $B(t,t')$ is independent of m.

Typically, to non-parametrically identify natural indirect effects, sequential ignorability requires us to account for all pre-treatment confounders affecting treatment, mediator, and outcome. Practically, researchers hardly observe and measure all confounders, and it is challenging to ensure all confounders are under control. Similar to causal inference, equivalently, this requires researchers fully understand or control how the treatment is assigned, which considerably narrows the scope of application.

3.2 Model-based Assumptions

Because of the challenge of non-parametric identification, in practice, researchers often rely on modeling assumptions to estimate the causal mediation effect. The traditional choice is the linear regression model (equation (2)-(5)). Several parametric assumptions are required to identify parameters and thus the indirect effect (MacKinnon 2012). In particular, we need two assumptions:

1. Correct function form, which primarily means linear in parameter and additivity;

2. No omitted variable, especially error terms ϵ_j should not correlate across equations.

Those function-form assumptions can also be interpreted by counterfactual languages (See Jo 2008; Sobel 2008). Generally, they correspond to unconfoundedness assumptions and additional function form assumptions.

Under the aforementioned assumptions, we obtain two famous estimators by different combinations of regression models. The difference estimator, $\hat{\tau} - \hat{\delta}$ uses equations (2) and (5). On the other hand, the product estimator, $\hat{\tilde{\beta}}\hat{\gamma}$, uses equations (2) and (3). MacKinnon et al. (1995) shows if models are correctly specified, then two estimators coincide.

It is worth noting that linear structural models implicitly assume no interaction effect between the treatment and mediator, and thus $\delta(0) = \delta(1)$ and $\eta(1) = \eta(0)$. This is similar to the above "no interaction effect assumption" in the non-parametric identification. If heterogeneous effects exist, the product estimator is biased, but the difference estimator remains unbiased (Glynn 2012). This is because the average of the product does not equal the product of averages:

$$\mathbb{E}[\tilde{\beta}^i \gamma^i] = \mathbb{E}[\tilde{\beta}^i] \times \mathbb{E}[\gamma^i] + Cov(\tilde{\beta}^i, \gamma^i) \tag{8}$$

If the covariance is not zero, then the product estimator is biased. From equation (8), we notice that the strong constant effect assumption can be replaced by a weaker assumption: the effect of the treatment on the mediator (γ) is not correlated to the effect of the mediator on the outcome ($\tilde{\beta}$), i.e., $Cov(\tilde{\beta}^i, \gamma^i) = 0$. Many other modified regression methods have been proposed in the literature (see Hong 2015). For example, we can add the interaction term in outcome model (2): $Y = \alpha_1 + \delta T + \tilde{\beta} M + \theta TM + \varepsilon_1$ (Imai et al. 2010a; Preacher et al. 2007).

In summary, mediation analysis under the linear regression model still requires several "exogeneity" assumptions to identify parameters in regression equations ($\tilde{\beta}$ and γ

or τ and δ). As shown in the Figure 3, generally speaking, both non-parametric and model-based assumptions require controlling U_1 and U_2. However, in practice, it is almost impossible to measure and control all such confounders. In observational studies using the linear model assumption, this is equivalent to asking researchers to find multiple fancy identification strategies in one study, which is not an easy task. To address this practical problem, our new identification idea can let researchers simultaneously identify both causal and mediation effects.

3.3 Identification Strategy

The above identification assumptions derive basic requirements for mediation analysis; however, they do not tell us how to satisfy those assumptions. In other words, we need identification strategies (Angrist and Krueger 1999; Samii 2016).

As one of the most widely used econometric tools, instrumental variables (IV) has been proposed to help identify indirect effects. Frölich and Huber (2017) consider two independent IVs, one for the treatment and the other for the mediator. With two IVs, they specify required assumptions and propose estimators for non-parametric identification of the indirect effect. Rudolph et al. (2021) extend the results to two related IVs. They also consider the case of a single IV for the treatment. Unfortunately, in this case, they conclude that we still need to rely on the assumption of no unobserved confounders of the mediator-outcome relationship.

Increased familiarity with identification strategies has allowed applied researchers to randomize or locate as-if-random treatments to identify the causal effect of treatment on the outcome and mediator. However, it is insufficient to identify the causal mediation effect. The main obstacle is that the mediator is not ignorable. By observing Figure 3, an interesting idea to address U_2 is to treat the treatment T as an IV for the mediator M. Therefore, we only need one IV for the treatment (or treatment is randomly assigned).

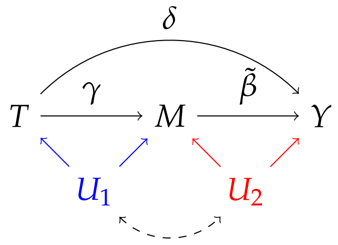

Figure 3: Mediation analysis generally requires addressing confounders U_1 and U_2.

To be a valid IV, however, it is well-known that T cannot have a direct effect on Y except from M. Sobel (2008) explores the identification of indirect effects under this exclusion assumption. It is clear that $\delta = 0$ is not a typical case. To account for the violation, Strezhnev et al. (2021) develops a useful sensitivity analysis method. Small (2011) proposes a different IV method to bypass the exclusion assumption; however, it requires the interaction between covariate X and a randomly assigned treatment T to be a valid IV for M. Recently, Dippel et al. (2019) found a new assumption allowing us only to use one IV for the treatment. The assumption adds constraints on the distribution of the unobserved confounding variables: unobserved confounding variables that jointly cause the treatment and the intermediate outcome are independent of the confounders that cause the intermediate and the final outcome, that is, U_1 and U_2 are independent.

Because most IV methods are developed under a linear structural model, one important feature is that they require a kind of "constant" effect assumption or zero covariance assumption ($Cov(\beta^i, \gamma^i)$) we mentioned before. [6] Other identification strategies exist, for example, experimental designs (Acharya et al. 2018), modified Difference-in-differences (DID), and synthetic control method (SCM) (See the survey by Celli (2022)).

In general, current identification strategies for mediation analysis still depend on several strong assumptions. Moreover, most are designed for IVs, which constrains the scope of application because it is not easy to find a good IV in many studies. Although our new

[6] See more detailed discussion by Hong (2015).

identification strategy also relies on some assumptions, we believe they are relatively easier to satisfy, and the strategy is not restricted to specific research designs.

4 Identification with Heterogeneous Effects

In this section, we introduce the new identification strategy, which starts with synthetic causal decomposition under the counterfactual approach but emphasizes the mechanical process as the structural approach. Under the new decomposition, we then convert the difficult mediation problem into a simple linear regression problem. It turns out to be quite general and simple to identify the causal mediation under this new structure. We then compare our new identification strategy with other methods and highlight that our method does not need the second part of ignorability assumptions or multiple IVs.

Recall that total causal effect can be decomposed into direct and indirect effects. We first consider the "no interaction effect" situation ($\delta = \delta(1) = \delta(0)$ and $\eta = \eta(1) = \eta(0)$). Because there is no interaction effect, the two decompositions in (1) are equivalent:

$$\tau = \mathbb{E}[Y^i(1, M^i(0)) - Y^i(0, M^i(0))] + \mathbb{E}[Y^i(1, M^i(1) - Y^i(1, M^i(0))] \tag{9}$$

$$= \mathbb{E}[Y^i(1, M^i(1)) - Y^i(0, M^i(1))] + \mathbb{E}[Y^i(0, M^i(1) - Y^i(0, M^i(0))] \tag{10}$$

$$= \mathbb{E}[Y^i(1, M^i(1)) - Y^i(0, M^i(1))] + \frac{\mathbb{E}[Y^i(0, M^i(1) - Y^i(0, M^i(0))]}{\mathbb{E}[M^i(1) - M^i(0)]} \times \mathbb{E}[M^i(1) - M^i(0)] \tag{11}$$

$$:= \delta + \beta\gamma \tag{12}$$

The first two lines are two decompositions. In the line (11),we multiply and divide the average indirect effect $\eta = \mathbb{E}[Y^i(0, M^i(1) - Y^i(0, M^i(0))]$ by the same term $\gamma = \mathbb{E}[M^i(1) - M^i(0)]$. It is the average effect of treatment on the mediator of interests. We define $\frac{\eta}{\gamma} = \frac{\mathbb{E}[Y^i(0,M^i(1)-Y^i(0,M^i(0))]}{\mathbb{E}[M^i(1)-M^i(0)]}$ by β, which denotes the ratio of how pure indirect effect changes according to one unit change of γ. Finally, we use simple notation to represent the final decomposition $\tau = \delta + \beta\gamma$.

In general, the parameter β can only be interpreted as the ratio representing how pure indirect effect η changes according to one unit change of γ. It cannot be interpreted as the effect of mediator M to outcome Y, as the $\tilde{\beta}$ as in model-based decomposition (5) and shown in Figure 3. However, in the linear SEM (2) and (3), $\tilde{\beta}$ and β are equivalent and can be interpreted as the average effect of mediator M to outcome Y. [7]

If we can identify β and γ, equivalently we can identify $\eta = \beta\gamma$. In most empirical studies, γ and τ are easy to identify if treatment is as if random through careful research designs. The remaining part is to identify the parameter β.

A critical insight in causal inference is the recognition that causal effects vary across populations and even among individuals. This implies that the γ is a random variable. Suppose we observe a random sample of γ (We will discuss how to get this sample in the section 5). Based on the above data-generating process, we also have a sample of τ, denoted as (τ_k, γ_k). Therefore, equation (12) can be written as $\tau_k = \delta + \beta\gamma_k$. Suppose δ is also random (with index k), then we obtain

$$\begin{aligned} \tau_k &= \delta_k + \beta\gamma_k && (14) \\ &= \mathbb{E}\delta_k + \beta\gamma_k + (\delta_k - \mathbb{E}\delta_k) && (15) \\ \Rightarrow \tau_k &= \mathbb{E}\delta_k + \beta\gamma_k + \varepsilon_k && (16) \end{aligned}$$

In the Line (15), we add and subtract the expectation of δ_k; in line (16), we define $\varepsilon_k = (\delta_k - \mathbb{E}\delta_k)$.

[7] Under linear SEM, we implicitly assume ε_1 and ε_2 are independent. From equation (2) and (3), we observe

$$\begin{aligned} \mathbb{E}[Y^i(0, M^i(1))] &= \alpha_1 + \tilde{\beta}(\alpha_2 + \gamma) \\ \mathbb{E}[Y^i(0, M^i(0))] &= \alpha_1 + \tilde{\beta}\alpha_2 \end{aligned} \qquad (13)$$

Therefore, $\eta = \mathbb{E}[Y^i(0, M^i(1))] - \mathbb{E}[Y^i(0, M^i(0))] = \tilde{\beta}\gamma$, where $\mathbb{E}[M^i(1) - M^i(0)] = (\alpha_2 + \gamma) - \alpha_2 = \gamma$.

Equation (16) should be familiar to readers: it is a simple linear regression model. To estimate β consistently (possibly β is also random, denoted by β_k), the key assumption is that the direct effect δ is uncorrelated with the effect of treatment on mediator γ. From now on, if a Greek letter has subscript k, then it denotes a random variable; otherwise, it is a constant.

Assumption 1 (No Correlation)**.** *$Cov(\gamma_k, \delta_k) = 0$.*

The assumption is equal to that of the traditional simple linear regression assumption $\mathbb{E}[\gamma_k \epsilon_k] = 0$ and implies $Cov(\gamma_k, \varepsilon_k) = 0$. To see this, $\mathbb{E}[\gamma_k \epsilon_k] = \mathbb{E}[\gamma_k(\delta_k - \mathbb{E}\delta_k)] = Cov(\gamma_k, \delta_k) = 0$ and $Cov(\gamma_k, \epsilon_k) = Cov(\gamma_k, \delta_k - \mathbb{E}\delta_k) = Cov(\gamma_k, \delta_k) = 0$. If there exist multiple mechanisms, generally, we should interpret δ_k as effects from all other possible mechanisms, see SI A and C. Therefore, the assumption requires that, γ, the treatment effect on the mediator of interest, is not correlated with the impacts of other mechanisms and direct effects. In scenarios without interaction effects between the treatment and the mediator, this assumption is typically easy to fulfill. In Proposition 1, we propose a simple estimator $\hat{\beta}$ to estimate the unknown β.

Proposition 1. *Let (τ, δ, γ) be random variables and as defined in* (11) *and* (12)*. Given the random sample $(\tau_k, \gamma_k)_{k \in K}$. Suppose $Var(\gamma_k) > 0$ and Assumption 1 holds.*

Consider estimator $\hat{\beta} = \frac{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)\tau_k}{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)^2}$.

(1) If β is a constant, then $\hat{\beta} \xrightarrow{p} \beta$ as $K \to \infty$;

(2) If β_k is a random variable, then $\hat{\beta} \xrightarrow{p} \mathbb{E}\beta_k$ as $K \to \infty$ under assumption $\beta_k \perp \gamma_k$ and thus η_k is consistently estimated by $\hat{\beta}\gamma_k$.

Proof. All proofs are provided in the SI. □

Estimator $\hat{\beta}$ is exactly the simple OLS estimator for the slope. The assumption $Var(\gamma_k) > 0$ is technical. In general, it guarantees that we have "random" observations of γ. The proposition says that if we assume the treatment effect on the mediator of interests is not correlated with the effects of other mechanisms, then we can consistently estimate β. Combined with the information of γ, we can estimate the average indirect effect. It's important to note that in other mediation analysis techniques, which do not account for heterogeneity, the estimates produced can actually be interpreted as average effects, especially in contexts where heterogeneity exists.

Remark 1 ($\beta_k \perp \gamma_k$)**.** *In the proposition 1, if β_k is a random variable, we need to assume $\beta_k \perp \gamma_k$ to ensure $\hat{\beta}$ is a consistent estimator of β. It may be a strong assumption. A simple solution is to use a model to approximate the correlation between β and γ. In other words, we could, for example, assume the following linear model $\beta = \theta_0 + \theta_1\gamma + \theta_2\gamma^2$. Then, replace it in the model (16) to be $\tau_k = \mathbb{E}\delta_k + (\theta_0 + \theta_1\gamma_k + \theta_2\gamma_k^2)\gamma_k + \varepsilon_k$. We can consistently estimate θ_0, θ_1, and θ_2 applying the usual linear regression estimator.*

Readers may be concerned that we are assuming a linear relationship between τ_k and γ_k. Actually, we do not assume such linearity; $\tau_k = \mathbb{E}\delta_k + \beta\gamma_k + \varepsilon_k$ is the structural model. There is an important distinction between this structural model and statistical linear regression models. First, in most cases, people assume the linear statistical relationship between data, that is, τ_k and γ_k here. Nevertheless, our model $\tau_k = \mathbb{E}\delta_k + \beta\gamma_k + \varepsilon_k$ is naturally guaranteed by the nature of the causal effect. In the counterfactual framework, we can always additively decompose the total causal effect into two pieces. Second, in statistical applications, people assume the expectation of the error term in their population model is zero: $\mathbb{E}\varepsilon_k = 0$. However, here, this property is guaranteed by construction, not by assumption: $\mathbb{E}\varepsilon_k = \mathbb{E}\delta_k - \mathbb{E}\delta_k = 0$. Because of this property, in contrast to OLS, the unbiasedness of our estimator requires a slightly weaker mean independence assumption ($\mathbb{E}[\delta_k|\gamma_k] = \mathbb{E}[\delta_k]$), rather than the zero conditional mean assumption ($\mathbb{E}[\delta_k|\gamma_k] = 0$). We summarize this result in SI D.

4.1 Interaction Effect

As mentioned early, under the assumption of no interaction effect between the treatment and mediator, identification assumption 1 is likely to hold in general. If we allow interaction effect, further justification may be required. However, actually, even if we allow interaction effect, if we are flexible to the parameter of interest (total or pure indirect effect), we can "purify" the interaction effect by using pure direct effect. Let me explain it by assuming a linear structural model with an interaction:

$$\mathbb{E}Y = \alpha_1 + \delta T + \tilde{\beta} M + \theta T M \tag{17}$$

$$\mathbb{E}M = \alpha_2 + \gamma T \tag{18}$$

Here, parameter θ captures the interaction effect between the treatment and mediator.

Recall that with an interaction effect, there are two ways to decompose total causal effect, and it is not hard to show them under the above linear structural model:

$$\begin{aligned}\tau_1 &= \delta(1) + \eta(0) \\ &= \mathbb{E}[Y^i(1, M^i(1)) - Y^i(0, M^i(1))] + \mathbb{E}[Y^i(0, M(1)) - Y^i(0, M(0))] \\ &= [\delta + \theta(\alpha_2 + \gamma)] + (\tilde{\beta}\gamma)\end{aligned} \tag{19}$$

and

$$\begin{aligned}\tau_2 &= \delta(0) + \eta(1) \\ &= \mathbb{E}[Y^i(1, M^i(0)) - Y^i(0, M^i(0))] + \mathbb{E}[Y^i(1, M(1)) - Y^i(1, M(0))] \\ &= (\delta + \theta\alpha_2) + [(\tilde{\beta} + \theta)\gamma]\end{aligned} \tag{20}$$

It needs to be noted that the total effect has a unique representation $\delta + \theta\alpha_2 + \tilde{\beta}\gamma + \theta\gamma$, i.e., $\tau_1 = \tau_2$. Therefore, relationship between τ and γ is unique. However, it has two interpretations $\delta(0) + \eta(1)$ and $\delta(1) + \eta(0)$ by considering different components at one time.

Corresponding to representation (16), in the decomposition (19), the "parameter" β is equal to $\tilde{\beta}$. We note that the total direct effect $\delta(1)$ contains γ. However, in the decomposition (20), the parameter $\beta = \tilde{\beta} + \theta$ and $\delta(0)$ does not contain γ. Obviously, in the latter decomposition, it is easy to have $Cov(\gamma_k, \delta_k) = 0$. Then, we can use OLS to estimate $\tilde{\beta} + \theta$ and the total indirect effect $\eta(1)$ in (20). However, in general, we cannot consistently estimate $\tilde{\beta}$ and thus pure indirect effect $\eta(0)$ with (19) because assumption 1 generally does not hold in this decomposition.

This example also highlights the interpretation of β, which is similar to the relationship between reduced-from and structural model. Although we estimate the same reduced-form model $\tau_k = \mathbb{E}\delta_k + \beta\gamma + \varepsilon_k$, under different assumptions, β represents different structural parameters. If there exists the interaction effect, β has two parts, one is $\tilde{\beta}$ (the effect of the mediator on the outcome), and the second part is θ (the interaction effect).

4.2 Comparison

What are the main advantages of our identification strategy? First, it does not require that mediator is ignorable. In other words, we allow unobserved confounders that simultaneously affect the mediator and the outcome variable (i.e., U_2 in the Figure 3). As mentioned in the section 3, current methods cannot efficiently address this unconfoundedness problem without further assumptions. Our methods bypass this problem by a novel decomposition and conversion (to a simple linear regression). Therefore, instead of collecting and controlling confounders, we only need researchers carefully scrutinize the research problem, the relationship among other mechanisms, the direct effect, and the mechanism of interest so that our identification assumption 1 holds.

Second, we allow researchers to simultaneously estimate both treatment and mediation effects. The causal mediation, is simply a byproduct after identifying the treatment effects (τ and γ). We do not need other advanced techniques to identify the indirect effect except simple OLS. We will introduce exact estimation methods and research designs in the next section. We can use both aggregate-level data and individual-level data to get the causal mediation effect. Therefore, we believe our methods can be applied in a variety of empirical studies.

5 Research Design Strategy

Our new identification strategy converts a challenging mediation analysis problem into a simple linear regression problem. The identification results, so far, assume that we already have a random sample of (γ, τ) at hand. However, practically, the question arises as to how to acquire this sample. In this section, we introduce two possible research designs: Heterogeneous Subgroup Design and Multiple Treatment Meta Design by exploring different sources of heterogeneity. Two modified estimators are also proposed to address the estimation uncertainty.

As we mentioned before, different values of treatment effects τ and γ are obtained from the heterogeneity in the population. We posit that two significant sources of heterogeneity exist. The initial source stems from the heterogeneous population receiving the treatment. Naturally, the same treatment may generate different (average) treatment effects for different population, for example, population in different country or area. To be concrete, for example, suppose researchers desire to understand how the treatment (mailing) affects turnout through social pressure in the Get-Out-The-Vote (GOTV) experiment (Gerber et al. 2008), as shown in the Figure 4. Formally, each individual is characterized by a vector of pre-treatment covariates $X = (X_1, X_2, ..., X_l)$ that can moderate treatment effects on the outcome and the mediator. We can subsequently define several subgroups

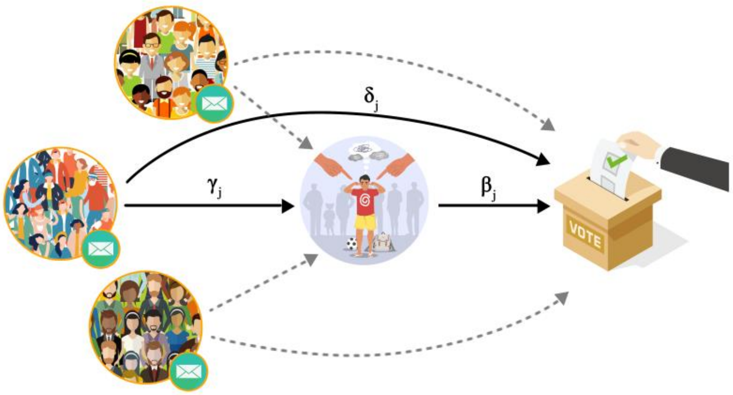

Figure 4: Heterogeneous Subgroup Design

G_k, where $k \in \{1, 2, ..., K\}$ according to X. Suppose X_1 is gender, and X_2 is age. We can define group $G_1 = \{X_1 = Male, X_2 > 30\}$, comprising individuals who are male and older than 30. Each individual i should belong to only one group. An assumption is that for each group, treatment generates different and independent average treatment effects τ and γ. How should we identify these groups? Relevant subgroups may be implied through theory where implications may arise. Alternatively, a data-driven method such as causal tree or forest can be used (Wager and Athey 2018).

Besides from variation in treatment effects across units, the second source of heterogeneity comes from the treatment. The key idea is that the treatment of interests has heterogeneous sub-types. For example, how does contact affect turnout? In the GOTV experiments, researchers conduct numerous experiments with heterogeneous treatments, including door-to-door canvassing, emails, phone calls, text messages, et al. All those treatments can be thought of as the sub-types of the meta-treatment: contact. See Figure 5. Formally, we still use $G_k \in \{T_1, T_2, ..., T_l\}$ to denote different sub-types of the treatment. If individual i belongs $G_k = T_j$, it means the individual receives treatment (type) T_j.

People may incorporate these two designs and define finer subgroups. The incorporated subgroup $G_k = \{T \in \mathbf{T}, X_1 \in \mathbf{X}_1, X_2 \in \mathbf{X}_2, ..., X_l \in \mathbf{X}_l\}$ is defined by treatment types and covariates, where $\mathbf{X}_l$ denotes a set possible values of X_l. For example, in the GOTV

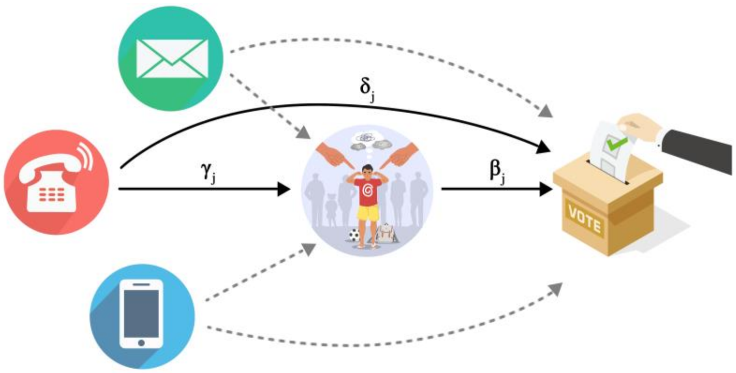

Figure 5: Multiple Treatment Meta Design

design, for each type of contact treatment, we can find subgroups defined by covariates. A possible subgroup could be $G_k = \{Phonecall, X_1 = Male, X_2 > 30\}$. If individual i is in this group, it implies that individual i is male, older than 30, and receive treatment phone call.

In all above designs, we can then estimate treatment effect $(\hat{\tau}_k, \hat{\gamma}_k)$ in each subgroup G_k, by using different identification strategies, including randomized controlled trials, IV, DID, SCM, regression discontinuity design, etc.

5.1 Aggregate Level Estimator

When we have those estimates, we cannot directly apply the simple OLS as we did in the proposition 1. The key reason is that $(\hat{\tau}_k, \hat{\gamma}_k)$ are only estimates; they are not the real (τ_k, γ_k). Therefore we treat data $(\hat{\tau}_k, \hat{\gamma}_k)$ as the noisy measurement. It is well-known that if the independent variable (γ_k here) is measured with error, it may lead to inconsistency.Particularly in the context of classical errors-in-variables (CEV) models, the estimate is prone to attenuation. Therefore, overlooking this aspect is not advisable, as it can significantly impact the accuracy of the results.

We now introduce two estimators. The first estimator only uses aggregate level data, similar to meta-analysis, where we only need estimated treatment effects $\hat{\tau}_k$ and $\hat{\gamma}_k$ rather than individual data. Therefore, the estimator can enable us to incorporate results from multiple studies (similar to the meta-analysis).

Most estimators in applied research are asymptotically normal. Therefore, without loss of generality, we also assume our estimates $(\hat{\tau}_k, \hat{\gamma}_k)$ are normally distributed around the true values (τ_k, γ_k).

Assumption 2 (Heterogeneous Measurement)**.**

$$\hat{\gamma}_k = \gamma_k + u_k \tag{21}$$

$$\hat{\tau}_k = \tau_k + v_k \tag{22}$$

where $u_k \sim N(0, \sigma^2_{uk})$ *and* $v_k \sim N(0, \sigma^2_{vk})$*,* $Cov(\gamma_k, u_k) = 0$*,* $Cov(\gamma_k, v_k) = 0$*,* $\sigma^2_{uk} > 0$*, and* $\sigma^2_{vk} > 0$*.*

In this assumption, as the classical setting, we also assume $Cov(\gamma_k, u_k) = 0$, and $Cov(\gamma_k, v_k) = 0$. However, departing from the CEV, we allow each estimate to have its own variance $\sigma^2_{\cdot k}$. This is more general, and more realistic because it is implausible that treatment effect has the same asymptotic variance across subgroups. Notably, our estimators introduced later are robust to the correlation between u_k and v_k. The following proposition shows that under this heterogeneous measurement assumption and identification assumption 1, the OLS $\hat{\beta}$ is still attenuated by $\lambda < 1$.

Proposition 2. *Suppose* (τ_k, γ_k) *satisfies the decomposition* (16)*:* $\tau_k = \mathbb{E}\delta_k + \beta_k \gamma_k + \varepsilon_k$*, and the observed random sample* $(\hat{\tau}_k, \hat{\gamma}_k)$ *follows the measurement assumption 2.*

Let σ^2_γ *be* $Var(\gamma_k)$ *and* $\lambda = \frac{\sigma^2_\gamma}{\sigma^2_\gamma + \sigma^2_{uk}}$*. Considering the estimator* $\hat{\beta} = \frac{\sum_{k=1}^{K} (\hat{\gamma}_k - \overline{\hat{\gamma}_k}) \hat{\tau}_k}{\sum_{k=1}^{K} (\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2}$*, under assumption* $\sum_{k=1}^{\infty} \frac{Var(\hat{\gamma}_k^2)}{k^2} < \infty$ *and assumption 1,*

(1) If β *is a constant, then* $\lambda^{-1}\hat{\beta} \xrightarrow{p} \beta$ *as* $K \to \infty$*;*

(2) If β_k *is a random variable, then* $\lambda^{-1}\hat{\beta} \xrightarrow{p} \mathbb{E}\beta_k$ *as* $K \to \infty$ *under assumption* $\beta_k \perp\!\!\!\perp \gamma_k$*.*

In the above proposition 2, we also assume $\sum_{k=1}^{\infty} \frac{Var(\hat{\gamma}_k^2)}{k^2} < \infty$. This technical assumption is required because, in the proof, we apply Kolmogorov's strong law of large numbers with independent but not identically distributed samples.

The proposition suggests using $\lambda^{-1}\hat{\beta}$ as a consistent estimator. For $\lambda = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \overline{\sigma_{uk}^2}}$, the numerator is the variance of true γ_k; in the denominator, $\overline{\sigma_{uk}^2}$ is the mean of the variance of $\hat{\gamma}_k$. Because the denominator is always larger than the numerator, $\lambda < 1$. In practice, we have data σ_{uk}^2 and therefore can calculate the sample average $\overline{\sigma_{uk}^2}$. However, we need an estimate of the unknown σ_γ^2, the variance of the true γ. The variance can be regarded as the "inter-study variance" in the random-effects model [8]. Many estimators in the meta-analysis literature exist (DerSimonian and Kacker 2007; DerSimonian and Laird 1986; Paule and Mandel 1982; Viechtbauer 2005).

Practically, we do not recommend using the previous estimator although it is useful to understand the attenuation (see also Bowden et al. 2016). We recommend the following Bivariate Correlated Errors and intrinsic Scatter (BCES) estimator (Akritas and Bershady 1996) and a simulation-extrapolation estimator (SIMEX) (Cook and Stefanski 1994). BCES estimator is widely used in Astrophysics and enables u_k and v_k to be correlated. The derivation uses the same logic we used in the proof of the proposition (see more discussion in the SI F). The estimator is

$$\hat{\beta}_{BCES} = \frac{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k}{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2 - \sum_{k=1}^{K}\sigma_{uk}^2} \tag{23}$$

[8] In the random-effects model, observed treatment effect y_i is assumed to be a function of the true treatment effect for the study θ_i and the sampling error e_i: $y_i = \theta_i + e_i$; and θ_i can be decomposed as $\mu + \delta_i$ where μ is the overall treatment effect and δ_i is the deviation of the i's-study's effect from the overall effect. The variance of δ_i is the inter-study variance. If, in the special case, it equals 0, we have the fixed-effect model.

with the asymptotic variance

$$\hat{\sigma}^2_{\hat{\beta}} = \frac{1}{K}\sum_{k=1}^{K}(\hat{\xi}_k - \overline{\hat{\xi}_k})^2 \tag{24}$$

where $\hat{\xi}_k = \frac{(\hat{\gamma}_k - \overline{\hat{\gamma}_k})(\hat{\tau}_k - \hat{\beta}\gamma_k - \hat{\delta}) + \hat{\beta}\sigma^2_{uk}}{\sigma^2_{uk} - \overline{\sigma^2_{uk}}}$, and $\hat{\delta} = \overline{\hat{\tau}_k} - \hat{\beta}_{BCES}\overline{\hat{\gamma}_k}$. Practically, in a finite sample, we seldom observe a large number of heterogeneous treatment effects so that k is large enough to satisfy the asymptotic requirement. Therefore, we propose using asymptotic refined percentile-t restricted wild (and pairs) bootstrap method to conduct statistical inference. See details in the SI G.

For the SIMEX, it uses a simple idea that the estimator for β can be regarded as a function of the variance of the measurement error, say $g(\sigma^2_{uk})$. The consistent estimator will be $g(0)$. The SIMEX method is to approximate the function g and extrapolate back to the case of no measurement error, $\sigma^2_{uk} = 0$. We will show both estimates in the simulation and application section 6, and wrap them in the R package.

5.2 Individual Level Estimator

Recall we use $G_k = \{T \in \mathbf{T}, X_1 \in \mathbf{X}_1, X_2 \in \mathbf{X}_2, ..., X_l \in \mathbf{X}_l\}$ to denote potential heterogeneous groups that are defined by treatment types and covariates. It is assumed that, in each G_k, average treatment effects $(\gamma_k, \tau_k, \delta_k)$ are different from each other. Let χ_{G_k} be the indicator function; $\chi_{G_k} = 1$ if individual i is in the group G_k and 0 if not.

For the individual-level estimator, we consider the linear model. For each group G_k, we assume different averages γ_k and δ_k and thus generate different τ_k. The model has the following form:

$$M^i = \sum_i \gamma_k T^i \chi_{G_k} + U^i \tag{25}$$

$$Y^i = \sum_i \delta_k T^i \chi_{G_k} + \beta M^i + V^i \tag{26}$$

where U^i and V^i denote all unobserved variables. Note that we allow U^i and V^i to be arbitrarily correlated so that M is "endogenous." Because M is "endogenous," we cannot use OLS to estimate parameter β.

To estimate β, an interesting idea is to use the treatment T_i be the IV for the mediator M. However, T is not a valid IV because it has direct effects δ on the outcome that are not mediated by M. If we assume treatment T is randomly assigned, this coincides with the imperfect IV problem considered by Kolesár et al. (2015). They consider several estimators and find that the estimator suggested by Anatolyev and Gospodinov (2011) is consistent, surprisingly, under the identification assumption 1. To see why models (25) and (26) are related to the non-zero correlation identification assumption, we replace M^i in the (26):

$$Y^i = \sum_i (\delta_k + \beta\gamma_k) T^i \chi_{G_k} + (\beta U^i + V^i) \tag{27}$$

Therefore, by regressing Y^i on T^i we obtain $\tau_k := \delta_k + \beta\gamma_k$, where γ_k can also be consistently estimated from (25) by regressing M^i on T^i. Subsequently, we face the same simple linear regression problem as the Proposition 1. Therefore, to consistently estimate β, our identification Assumption 1 is required.

In the observational study, T is not randomly assigned. If the causal identification strategy is IV, then we can simply add one more equation:

$$T^i = \theta Z^i + e_i \tag{28}$$

where Z is the IV for the treatment T.

6 Simulation and Application

In this section, we employ Monte Carlo simulations to evaluate the effectiveness of the BCES and SIMEX estimators, with a focus on scenarios involving small sample sizes. This focus is pragmatic, as identifying a large number of subgroups is often challenging in actual research conditions. Furthermore, we apply our methodology to real data from two distinct experiments—one using aggregate data and the other using individual-level data—to illustrate its application in real studies.

6.1 Simulation

In the simulation, we posit that the decomposition takes the form: $\tau_k = 4 + \beta\gamma_k + N(0, 1)$. We sample the true parameter γ from a normal distribution with a mean of 2 and a standard deviation of 1. The (initial) standard errors for τ and γ are sampled from a gamma distribution where both the shape and rate parameters are set to 1. The pseudo-code is shown below.

We begin by setting $\beta = 0$ and compare the result with the theoretical rejection rate—or size—of 0.05 in a two-tailed test. The left part of the table 1 shows the simulation results. In general, the BCES estimator with asymptotic variance has the poorest performance, exhibiting a rejection rate of 0.422 even with a sample size of 5. BCES with non-parametric pairs bootstrap moderately reduce the over rejection in large sample size. The BCES esti-

Algorithm 1: Monte Carlo Simulation

```
**Input:** : $k > 0, \beta, \delta = 4$
$\gamma \sim Norm(2,1), \epsilon \sim Norm(0,1),$
$\tau_k = \delta_k + \beta\gamma_k + \epsilon$
$\sigma_{uk} \sim Gamma(1,1), \sigma_{vk} \sim Gamma(1,1)$
Initialization: $b = 1$
**repeat**
    $u_k \leftarrow Norm(0, \sigma_{uk}), v_k \leftarrow Norm(0, \sigma_{vk})$
    $\hat{\gamma}_k \leftarrow \gamma_k + u_k, \hat{\tau}_k \leftarrow \tau_k + v_k$
    get $\hat{\beta}_b$ from BCES and SIMEX
    $k \leftarrow k + 1$
**until** $b = B$
**Output:** $\{\hat{\beta}_b\}$
```

mator, when adjusted with a non-parametric pairs bootstrap, shows a moderate decrease in over-rejection for larger sample sizes. However, both the BCES with a restricted wild bootstrap and the SIMEX estimator significantly improve performance with small sample sizes. Notably, the empirical rejection rate for SIMEX closely matches the theoretical rate (0.05) even with an extremely small sample size of 5 observations. This indicates that the original BCES estimator should not be used without adjustments.

We next evaluate the statistical power of the estimators by setting $\beta = 2$, which yields an average indirect effect constituting approximately 50 percent of the total treatment effect—a reasonable benchmark. The BCES estimator with pairs bootstrap demonstrates higher power with small sample sizes, whereas the SIMEX and restricted wild bootstrap BCES exhibit considerable power when the sample size exceeds 10. Overall, SIMEX consistently outperforms the others in our simulations.

K	Rejection Rate: $\beta = 0$				Rejection Rate: $AIE = 50\%$ Total Effect		
	BCES	BCES (pairs)	BCES (wild)	SIMEX	BCES (pairs)	BCES(wild)	SIMEX
100	0.034	0.042	0.038	0.092	0.192	0.548	0.988
50	0.067	0.082	0.04	0.09	0.16	0.302	0.854
30	0.166	0.002	0.05	0.066	0.024	0.438	0.926
10	0.338	0.228	0.05	0.084	0.384	0.166	0.27
5	0.422	0.184	0.018	0.058	0.222	0.102	0.168

Table 1: Monte Carlo Simulation and Rejection rate.

	Brazil	China	Costa Rica	Liberia	Peru	Uganda
Resource	Groundwater	Surface water	Groundwater	Forest	Forest	Forest
Components of treatment						
Community workshops	✓	-	✓	✓	✓	✓
Monitor selection, training,incentives	✓	✓	✓	✓	✓	✓
Monitoring of the resource	✓	✓	✓	✓	✓	✓
Dissemination to citizens	✓	✓	✓	✓	✓	✓
Dissemination to management bodies	-	(alternative arm)	✓	✓*	✓*	✓*

Table 2: Summary Table from Slough et al. (2021). * In the forest studies, the community constitutes at least one of the possibly overlapping management bodies.

6.2 Application I: Governance on Resources (Aggregate level Data)

Evidence in Governance and Politics (EGAP) [9] funds and coordinates multiple field experiments on different topics across countries. This collaborative research model is called "Metaketa Initiative." In Metaketa III, they examine the effect of community monitoring on common pool resources (CPR) governance. To causally answer this question, Slough et al. (2021) conducted six harmonized experiments with the same 'meta' treatment (community monitoring) but heterogeneous CPRs and treatment sub-types, as shown in the table 2.

In their study, the authors report effects on multiple outcome variables, which include resource use, user satisfaction, user knowledge about community's CPRs, and resource stewardship. They also investigate the underlying mechanism: how monitoring affects those outcomes through different channels. However, their analysis is limited to examining the treatment effects on mediators, which does not necessarily delineate the precise

[9] https://egap.org/

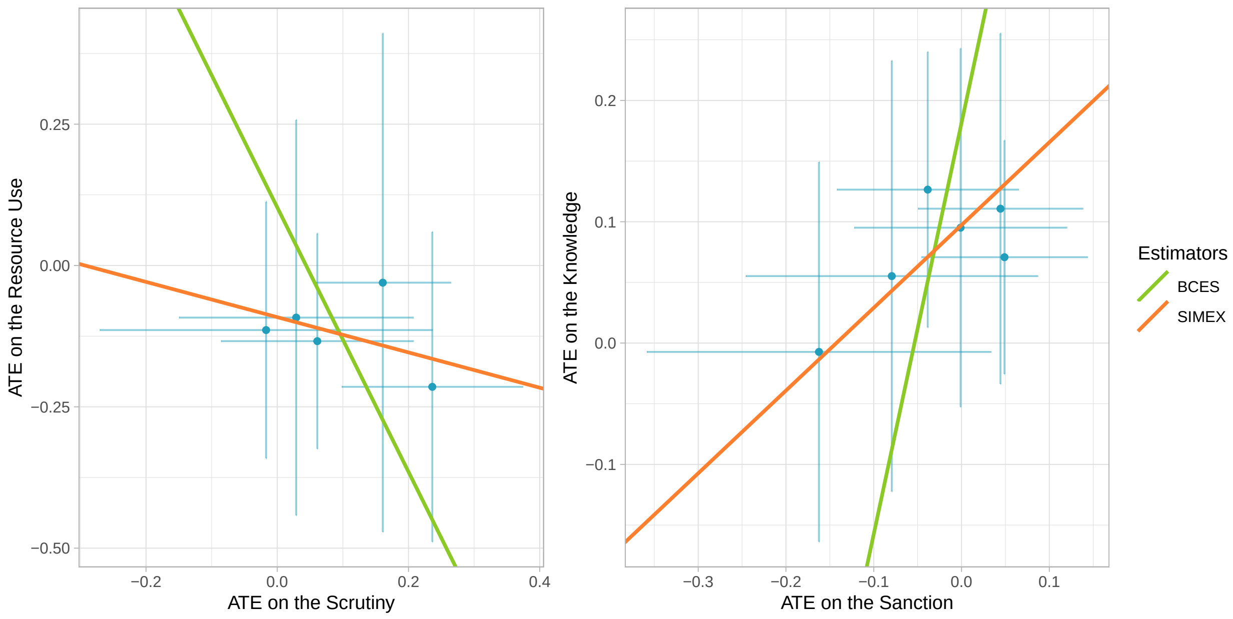

Figure 6: Heterogeneous Subgroup Design

causal mediation effects. We intend to use their dataset to accurately quantify the mediation effect. Specifically, we aim to address two questions regarding the mechanisms:

(1) How does monitoring influence resource usage through increased user scrutiny of CPR management authorities?; and

(2) How does monitoring affect user knowledge about community CPRs through altered perceptions of sanction likelihood for CPR misuse?

The six experiments naturally provide us with six subgroups. While it's feasible for researchers to further segment subgroups within each experimental site, our focus will be on these six primary subgroups. This approach demonstrates that our estimators, namely BCES and SIMEX, are effective even with such limited sample sizes. For the application of these estimators, we require specific data: (1) the average treatment effects on both the outcome and the mediator, and (2) the standard errors associated with these effects. The estimated β_{bces} (pairs bootstrap), β_{simex} and corresponding data points are shown in figure 6.

In the left panel, the average effects on both the mediator (scrutiny) and the outcome (resource use) across five experiments are represented by blue points with error bars. It's important to note that in the Chinese experiment, the scrutiny measure wasn't gauged at the citizen level as in other sites, and therefore it is omitted from this analysis. Generally, these original estimates lack precision, which can be attributed to the moderately sample size in each experiment. To obtain the mediation effect, we need to estimate β. We find that $\beta_{bces} = -2.34$ with $p = 0.06$ (pairs bootstrap); however, SIMEX and restricted wild bootstrap BCES are not statistically significant at 0.1. It is not surprising because we already see pairs bootstrap BCES has higher power compared to SIMEX under this extremely small sample size. To interpret the result, we can simply multiply the β and γ to obtain the causal mediation effect. [10] For example, in Uganda, the average treatment effect on resource use is -0.09 and the average treatment effect on scrutiny is 0.03; thus the monitoring effect through scrutiny on CPR management authorities is -0.07, which accounts for 77.78% of the total average treatment effect.

Similarly, in the right panel, we illustrate the average treatment on the mediator (sanction) and the outcome (knowledge). Remarkably, even with six observations, two estimates are still significantly congruent on the positive sign ($\beta_{simex} = 0.67, p = 0.05$, and $\beta_{bces} = 3.39, p = 0.02(pairs)$). Therefore, for example, in China, the average treatment effect on the outcome knowledge is 0.11 and the average treatment effect on the mediator sanction is 0.04. This allows us to infer that the indirect effect, mediated through the likelihood of being sanctioned, accounts for 27% of the total effect of monitoring on enhancing knowledge about the community's common pool resources (CPR).

[10] We assume β is constant here. In practice, the validity of the assumption requires further justification and examination. The results presented here should be viewed primarily as illustrative.

6.3 Application II: Information Effect on Voting (Individual-level data)

Accountability is a cornerstone of democracies and is fundamental to good governance. However, in reality, voters often lack sufficient information about politicians' performance. Many organization and civil society groups have dedicated efforts to disseminate such information to the electorate. A pivotal question arises: "Do informational interventions influence voters' behaviors and thereby promote accountability? If yes, what is the key mechanism?" Numerous field and survey experiments have sought to quantify this treatment effect; yet the findings are inconclusive. (see Dunning et al. 2019b; Incerti 2020) Furthermore, our understanding of how information influences voting behavior is limited. In a few experiments, researchers have measured intermediate outcomes and explore potential mechanisms. Nevertheless, these intermediate outcomes are not ignorable given the treatment status, making the estimation of mediation effects challenging. In this section, we will illustrate the use of our method in one field experiment from Benin, demonstrating how to identify and estimate the causal mediation effect in an information experiment using individual-level data.

Around 2015 National Assembly elections in Benin, as a part of Metaketa I, Adida et al. (2019) randomly disseminated information about the performance of incumbent legislators to voters through videos. These videos provided official data on four key performance dimensions: (1) attendance rate at legislative sessions, (2) frequency of posing questions during these sessions, (3) committee attendance rate, and (4) productivity of committee work. One of the primary outcome variables was individual voting choice, which was captured via baseline and endline surveys. The surveys also gathered intermediate variables, such as voters' perceptions of the incumbents' effort/hardworking. Overall, the intervention did not significantly affect the incumbents' vote shares, aligning with the results from most other field experiments (Dunning et al. 2019b). However, a subsequent meta-analysis highlighted a notable correlation between voters' perceptions of effort and support for incumbents (Dunning et al. 2019a, p354). As emphasized by authors, this correlation does not illuminate any causal relationship due to the design. Nevertheless, it indicates a potential indirect effect of information on voting behavior mediated by perceptions of hard work. Thus, we intend to apply our method to their individual-level data to directly estimate this causal mediation effect.

To apply our method to individual-level data, the initial step involves identifying potential heterogeneous subgroups. This identification can be achieved theoretically or through data-driven methods. We employ the widely-used causal tree approach to detect these subgroups, estimating the heterogeneous treatment effect on the mediator (effort) using individual pre-treatment covariates, such as age, gender, wealth, and political attitudes.[11] As depicted in Figure 7, the informational effect on the perception of effort varies according to factors like wealth status, prior beliefs about effort, gender, and certain political attitudes.

Next, we estimate the average treatment effect on the vote choice across these seven subgroups. The estimates are then utilized to calculate the indirect effect using BCES or SIMEX, as aggregate-level data. The final results are illustrated in Figure 8. We found that the estimated β is 0.17, which is significant at the 0.1 level (according to the SIMEX analysis). Consequently, we deduce that, while the overall information effect may not be substantial enough to be detected in the field experiment, there is a significant indirect effect through the voters' perception of the politicians' effort. Specifically, if voters perceive that the politician has exerted more effort, based on the information they have received, they are more likely to vote for the incumbent.

[11] For details, refer to the replication code. Additional results are presented in Supplementary Information I.

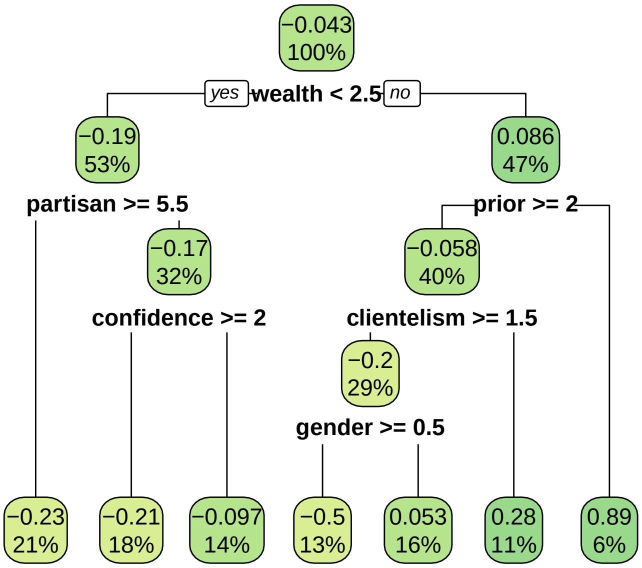

Figure 7: Subgroups detected by Causal trees

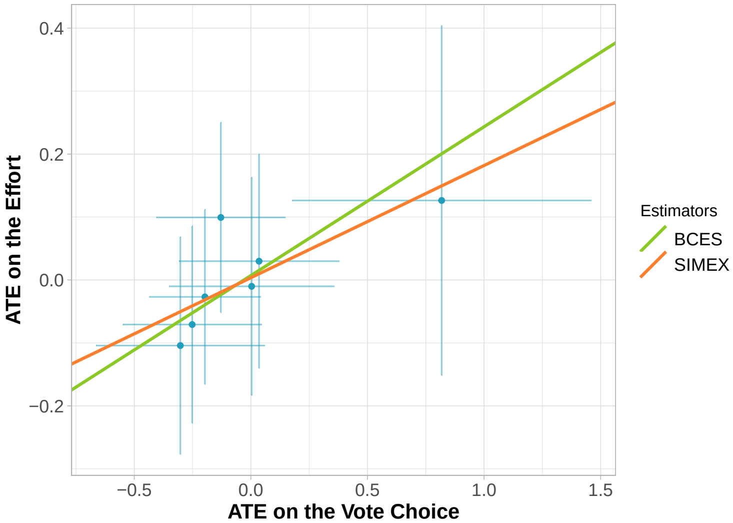

Figure 8: Heterogeneous Subgroup Design

7 Conclusion

Understanding the causal mechanism is essential for social science. Mediation analysis offers powerful statistical tools for quantifying the causal mediation effect. However, straightforward methods that do not rely on stringent sequential ignorability assumptions and can be readily applied across a range of empirical studies are still lacking.

In this study, we propose a novel identification assumption and strategy that can enable researchers easily estimate causal mediation effects. Within the potential outcome framework, we introduce a causal decomposition that emphasizes the mechanism process similar to the structural approach. This innovative decomposition converts the intricate mediation problem into a simple linear regression problem. Based on the novel zero correlation assumption, once researchers identify the treatment effects on the mediator and the outcome, our approach can consistently estimate the indirect effect. The method exploits the causal heterogeneity and proves to be remarkably simple.

While our method alleviates the need for stringent assumptions, it is important to acknowledge that it is not a universal solution for all mediation analysis scenarios. Firstly, in certain cases, our method is not entirely model-free; for instance, approximating the correlation between β_k and γ_k may be necessary. However, relying on parametric identification should not be seen as a drawback. As Hansen (2022, p56) notes, using such assumptions for identification is often considered a practical, albeit second-best, solution in modern econometrics. Secondly, there are several avenues for further exploration. Questions like extending the method to non-binary treatments, identifying necessary assumptions for correlated mechanisms, or integrating other causal identification strategies beyond instrumental variables (IV) for individual-level estimators remain open. Lastly, our method suggests a promising avenue to bridge causal mediation with causal moderation, indicating the potential for discovering other effective methodological combinations.

References

Acharya, Avidit et al. (2016). "Explaining causal findings without bias: Detecting and assessing direct effects". *American Political Science Review* 110.3, pp. 512–529.

— (2018). "Analyzing causal mechanisms in survey experiments". *Political Analysis* 26.4, pp. 357–378.

Adida, Claire et al. (2019). "Under what conditions does performance information influence voting behavior? Lessons from Benin". *Information, accountability, and cumulative learning: lessons from Metaketa I*. Cambridge University Press, pp. 81–117.

Akritas, Michael G and Matthew A Bershady (1996). "Linear regression for astronomical data with measurement errors and intrinsic scatter". *arXiv preprint astro-ph/9605002*.

Anatolyev, Stanislav and Nikolay Gospodinov (2011). "Specification testing in models with many instruments". *Econometric Theory* 27.2, pp. 427–441.

Angrist, Joshua D and Alan B Krueger (1999). "Empirical strategies in labor economics". *Handbook of labor economics*. Vol. 3. Elsevier, pp. 1277–1366.

Avin, Chen et al. (2005). "Identifiability of path-specific effects".

Baron, Reuben M and David 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, p. 1173.

Bowden, Jack et al. (2015). "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression". *International journal of epidemiology* 44.2, pp. 512–525.

Bowden, Jack et al. (2016). "Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I 2 statistic". *International journal of epidemiology* 45.6, pp. 1961–1974.

Celli, Viviana (2022). “Causal mediation analysis in economics: Objectives, assumptions, models”. *Journal of Economic Surveys* 36.1, pp. 214–234.

Cook, John R and Leonard A Stefanski (1994). “Simulation-extrapolation estimation in parametric measurement error models”. *Journal of the American Statistical association* 89.428, pp. 1314–1328.

DerSimonian, Rebecca and Raghu Kacker (2007). “Random-effects model for meta-analysis of clinical trials: an update”. *Contemporary clinical trials* 28.2, pp. 105–114.

DerSimonian, Rebecca and Nan Laird (1986). “Meta-analysis in clinical trials”. *Controlled clinical trials* 7.3, pp. 177–188.

Dippel, Christian et al. (2019). “Mediation analysis in IV settings with a single instrument”. *Working Paper*.

Dunning, Thad et al. (2019a). *Information, accountability, and cumulative learning: Lessons from Metaketa I*. Cambridge University Press.

Dunning, Thad et al. (2019b). “Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials”. *Science advances* 5.7, eaaw2612.

Frölich, Markus and Martin Huber (2017). “Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables”. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 1645–1666.

Fu, Jiawei and Tara Slough (2023). “Heterogeneous Treatment Effects and Causal Mechanisms”. *Working Paper*.

Gallop, Robert et al. (2009). “Mediation analysis with principal stratification”. *Statistics in medicine* 28.7, pp. 1108–1130.

Gerber, Alan S et al. (2008). “Social pressure and voter turnout: Evidence from a large-scale field experiment”. *American political Science review* 102.1, pp. 33–48.

Glynn, Adam N (2012). “The product and difference fallacies for indirect effects”. *American Journal of Political Science* 56.1, pp. 257–269.

Hafeman, Danella M and Tyler J VanderWeele (2011). “Alternative assumptions for the identification of direct and indirect effects”. *Epidemiology*, pp. 753–764.

Hansen, Bruce (2022). *Econometrics*. Princeton University Press.

Holland, Paul W (1986). “Statistics and causal inference”. *Journal of the American statistical Association* 81.396, pp. 945–960.

Hong, Guanglei (2015). *Causality in a social world: Moderation, mediation and spill-over*. John Wiley & Sons.

Imai, Kosuke et al. (2010a). “A general approach to causal mediation analysis.” *Psychological methods* 15.4, p. 309.

Imai, Kosuke et al. (2010b). “Identification, inference and sensitivity analysis for causal mediation effects”.

Imai, Kosuke et al. (2011). “Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies”. *American Political Science Review* 105.4, pp. 765–789.

Imbens, Guido W (2020). “Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics”. *Journal of Economic Literature* 58.4, pp. 1129–1179.

Incerti, Trevor (2020). “Corruption information and vote share: A meta-analysis and lessons for experimental design”. *American Political Science Review* 114.3, pp. 761–774.

Jo, Booil (2008). “Causal inference in randomized experiments with mediational processes.” *Psychological Methods* 13.4, p. 314.

Kolesár, Michal et al. (2015). “Identification and inference with many invalid instruments”. *Journal of Business & Economic Statistics* 33.4, pp. 474–484.

MacKinnon, David P (2012). *Introduction to statistical mediation analysis*. Routledge.

MacKinnon, David P et al. (1995). "A simulation study of mediated effect measures". *Multivariate behavioral research* 30.1, pp. 41–62.

Paule, Robert C and John Mandel (1982). "Consensus values and weighting factors". *Journal of research of the National Bureau of Standards* 87.5, p. 377.

Pearl, J (2001). "Direct and indirect effects Paper presented at: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence".

Pearl, Judea (2014). "Interpretation and identification of causal mediation." *Psychological methods* 19.4, p. 459.

Preacher, Kristopher J et al. (2007). "Addressing moderated mediation hypotheses: Theory, methods, and prescriptions". *Multivariate behavioral research* 42.1, pp. 185–227.

Robins, James M (2003). "Semantics of causal DAG models and the identification of direct and indirect effects". *Oxford Statistical Science Series*, pp. 70–82.

Robins, James M and Sander Greenland (1992). "Identifiability and exchangeability for direct and indirect effects". *Epidemiology* 3.2, pp. 143–155.

Rudolph, Kara E et al. (2021). "Causal mediation with instrumental variables". *arXiv preprint arXiv:2112.13898*.

Samii, Cyrus (2016). "Causal empiricism in quantitative research". *The Journal of Politics* 78.3, pp. 941–955.

Slough, Tara et al. (2021). "Adoption of community monitoring improves common pool resource management across contexts". *Proceedings of the National Academy of Sciences* 118.29, e2015367118.

Small, Dylan S (2011). "Mediation analysis without sequential ignorability: using baseline covariates interacted with random assignment as instrumental variables". *Journal of Statistical Research* 46.2, pp. 91–103.

Sobel, Michael E (2008). "Identification of causal parameters in randomized studies with mediating variables". *Journal of Educational and Behavioral Statistics* 33.2, pp. 230–251.

Strezhnev, Anton et al. (2021). “Testing for Negative Spillovers: Is Promoting Human Rights Really Part of the “Problem”?” *International Organization* 75.1, pp. 71–102.

Ten Have, Thomas R and Marshall M Joffe (2012). “A review of causal estimation of effects in mediation analyses”. *Statistical Methods in Medical Research* 21.1, pp. 77–107.

VanderWeele, Tyler (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, Tyler J (2011). “Principal stratification–uses and limitations”. *The international journal of biostatistics* 7.1, p. 0000102202155746791329.

— (2013). “A three-way decomposition of a total effect into direct, indirect, and interactive effects”. *Epidemiology (Cambridge, Mass.)* 24.2, p. 224.

Viechtbauer, Wolfgang (2005). “Bias and efficiency of meta-analytic variance estimators in the random-effects model”. *Journal of Educational and Behavioral Statistics* 30.3, pp. 261–293.

Wager, Stefan and Susan Athey (2018). “Estimation and inference of heterogeneous treatment effects using random forests”. *Journal of the American Statistical Association* 113.523, pp. 1228–1242.