Identification Strategy for Mediation Analysis Relying on Heterogeneous Effects

Jiawei Fu[*]

March 6, 2024

Abstract

Understanding causal mechanisms is indispensable for explaining and generalizing the empirical phenomenon. Causal mediation analysis provides statistical techniques to quantify mediation effects. However, current methods require strong identification assumptions or demanding research designs. To make it easier to use, we have developed a new identification strategy to convert the difficult mediation problem into a simple linear regression problem using a novel decomposition that combines counterfactual and structural frameworks. The new method establishes a novel link between causal mediation and causal moderation; Notably, it demonstrates relative simplicity than conventional methods. Several research designs and estimators are discussed and proposed in the study making the identification strategy easily applicable to various empirical studies. Overall, the new identification strategy offers a new direction for exploring causal mediation and causal moderation.

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

[*] 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, applying mediation analysis frequently encounters hindrances owing to onerous assumptions. This study proposes a novel and straightforward identification strategy to facilitate the conduction of causal mediation analysis. In general, our strategy can convert a difficult causality problem into a simple data problem, which allow researchers estimate causal mediation effect and treatment effect simultaneously without more identification burdens.

Figure 1 illustrates some basic concepts of mediation analysis. Because of the existence of a mediator, mediation analysis requires stricter assumptions than usual causal inference. 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), while mediation analysis requires the ignorability of the mediator (U_2 and U_4). In a typical empirical study, satisfying and justifying these two identification problems is challenging. Therefore, a useful and successful causal mediation identification strategy should simultaneously estimate the treatment and mediation effects, or at least not impose additional identification burdens. Through our demonstration, 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 analyzing mediation in social science. Inspired by Baron and Kenny (1986), scholars typically employ linear regression models and utilize the "product" or

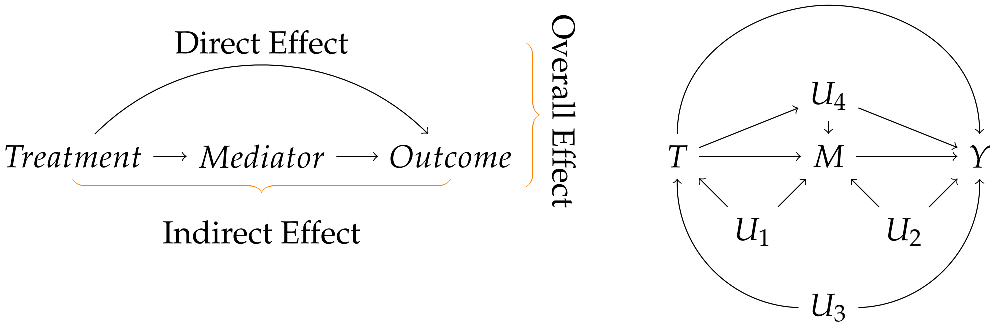

Figure 1: The left panel illustrates the basic decomposition of the overall treatment effect. The right panel is one example of the directed acyclic graph. T is the treatment variable, M is the mediator, Y is the outcome variable, and $U_j, j = 1, 2, 3, 4$ are confounders.

"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 (Pearl 2001). By decomposing the overall treatment effect into natural direct and indirect effects, this approach deepens our understanding of the causal mechanism and clarifies unconfounding 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 is also needed to be ignorable, given the actual treatment status. The second ignorability requirement is the primary obstacle to application.

Our novel identification strategy incorporates the strength of two approaches and does not require strong exogeneity or parametric assumption. Under the counterfactual approach, we first deconstruct the natural indirect effect by highlighting the treatment effect on the mediator. After that, we do not directly identify the natural indirect effect; instead, as we show in section 4, we convert it to be a simple linear regression problem and aim to identify the slope coefficient. If the slope is identified, we are able to identify the causal mediation effect. In the converted simple linear regression, the dependent variable is the average treatment effect on the outcome, and the independent variable is the average treatment effect on the mediator. In other words, these two effects are treated as our observed "data". As OLS, the critical identification assumption is that the average treatment effect on the mediator is not correlated to other mechanisms (corresponding to the error term). Under the 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, people have explored it in the problem of invalid instrumental variables (Kolesár et al. 2015) and Mendelian Randomization (Bowden et al. 2015). Nevertheless, this is the inaugural instance of its rediscovery 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 frost (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 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 the meta-analysis. The second class is the individual-level estimator, which can provide more precise results by assuming some parametric structures. In general, as long as researchers can identify treatment effects, which is quite simple given that there exits so many causal identification strategies, researchers can easily identify causal mediation using our strategy under mild assumptions.

Our identification strategy turns out to be extremely simple, which does not need any advanced techniques other than simple linear regression. One reason is that it exploits the casual heterogeneity, which is commonly ignored in the 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 really talk to each other in theory. 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) firstly 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. In this section, we introduce and compare the counterfactual and structural approaches, if necessary, with the help of directed acyclic graph (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 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 $t' = 0$

[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.

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 that they would be when receiving the 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 multiple mechanisms, apart from the convention, we use $j-$ 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]:

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

Similar to the fundamental problem of causal inference, we are typically interested in the average level. 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

[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.

discussed in supplement 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 $\eta^i(t) = Y^i(t, M^i(1)) - Y^i(t, M^i(0))$, if there exists interaction effect, the value of $\delta^i(t)$ and $\eta^i(t)$ may depend on t. 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.

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 the basic structure we refer to.

If people do not assume "no interaction effect", we can further decompose the total

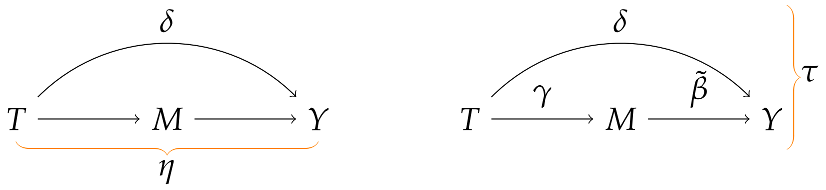

Figure 2: Decompositions with Counterfactual and Structural Approaches

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, 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 the potential DGP.

Unlike the counterfactual approach, structural models implicitly incorporate a few parametric assumptions, such as the linear relationship 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 quite 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, $\eta(t) = Y(t, M(1)) - Y(t, M(0))$ incorporates two effects into one formula, which obscures the mechanical pathways of how the treatment is mediated by intermediate variables. As subsequently demonstrated, re-introduce γ into the counterfactual decomposition may lead to new insights about the mediation analysis.

3 Identification Assumption

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.

3.1 Non-parametric Assumptions

To non-parametrically identify the causal mediation effect, we require the sequential ignorability assumption. In contrast to the causal inference literature, multiple versions of sequential ignorability assumption exist [4]. One of the most concise ones 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:

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

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 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, which makes it challenging to interpret and cannot be satisfied by any experimental design because we can not let the same

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

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 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 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 difficulty of non-parametric identification, in practice, researchers mainly rely on modeling assumptions to estimate the causal mediation effect. The traditional choice is the linear regression model (equation (2)-(5)). Basically, as with all regression analyses, 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, $\hat{\hat{\beta}}\hat{\gamma}$, 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 those confounders. In observational studies with 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, or at least not add more identification burdens.

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

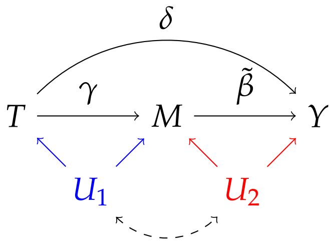

Figure 3: Mediation analysis generally requires addressing confounders U_1 and U_2.

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.

Evidently, owing to the prevalence of causal identification strategies, researchers are good at making the treatment assignment as if random and thus identify the causal effect of treatment on the outcome and mediator. However, it is insufficient to identify the causal mediation effect even if we assume linear regression models. 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). 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 [5]. Other identification strategies exist, for example, modified DID, SCM, and experiment designs (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 daily application. Although our new identification strategy also relies on some assumptions, we believe they are easy 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

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

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 (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, parameter β can only be interpreted as the ratio representing how pure indirect effect η changes according to one unit change of γ. It can not 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. [6]

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

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

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

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

designs. The remaining part is to identify the parameter β.

Now, suppose γ is random, and 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). If this is the case, 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 $\varepsilon_k = (\delta_k - \mathbb{E}\delta_k)$.

I hope the equation (16) arouses memories of the old: 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 $Cov(\gamma_k, \varepsilon_k) = Cov(\gamma_k, \delta_k - \mathbb{E}\delta_k) = Cov(\gamma_k, \delta_k) = 0$. If there exist multiple mechanisms, generally, we should interpret δ_k as all possible mechanisms other than the mechanism (mediator) of interests, see SI A and C. Therefore, the assumption requires that γ the treatment effect on the mediator of interest is not correlated with all other mechanisms and direct effects. Under "no interaction effect between treatment and mediator", this assumption is generally easy to satisfy.

In the next 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 hold.*

Considering 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. Assumption $Var(\gamma_k) > 0$ is technique. 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 other mechanisms, then we can consistently estimate β. Combined with the information of γ, we can estimate the average indirect effect.

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 consistent. 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.*

People may mistakenly think that we "assume" a linear regression model. Actually, $\tau_k = \mathbb{E}\delta_k + \beta\gamma_k + \varepsilon_k$ is the structural model. A huge difference exists between this structural model and statistical linear regression models people use daily. 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 the result in the SI D.

As mentioned early, under "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 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 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 pointed out 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 find 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 hardly holds 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).

4.1 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 the 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.

Generally, different values of treatment effects τ and γ must be obtained from the heterogeneity in the population. We posit that two significant sources of heterogeneity exist.

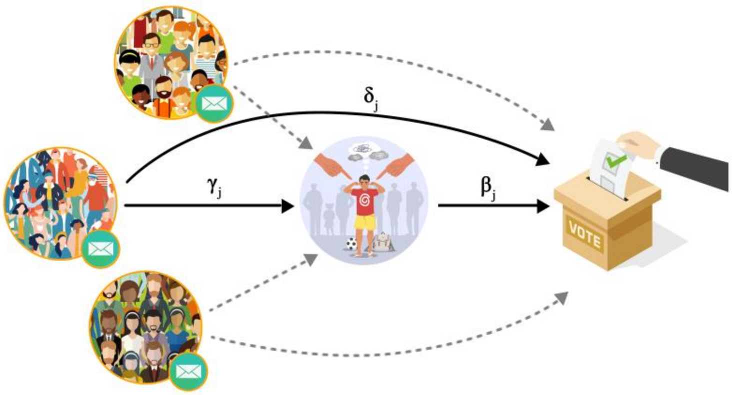

Figure 4: Heterogeneous Subgroup Design

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 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 to identify these groups? It can be implied through theory where implications may arise. Alternatively, a data-driven method such as causal tree or forest (Wager and Athey 2018).

Besides from receivers, 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 treatment, including door-to-door canvass, email,

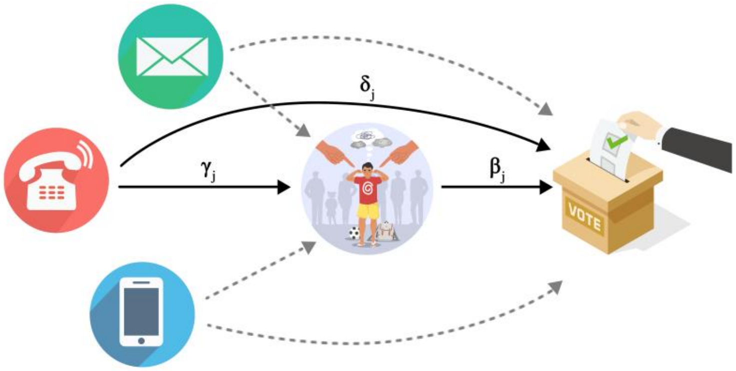

Figure 5: Multiple Treatment Meta Design

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 $G_k = \{T_1, T_2, ..., T_l\}$ 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 $G_k = \{T, X_1, X_2, ..., X_l\}$ is defined by treatment types and covariates. For example, in the GOTV 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.

We use S to denote any strategy that can aid us to identify treatment effects, including randomized controlled trials (RCT), IV, DID, RD, SCM, etc. For heterogeneous subgroup design, we can use S to estimate treatment effect $(\hat{\tau}_k, \hat{\gamma}_k)$ in each subgroup G_k. For Multiple Treatment Meta Design, each treatment type, we can also obtain the required $(\hat{\tau}_k, \hat{\gamma}_k)$ using different S.

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. Especially in the classical errors-in-variables (CEV), the estimate will be attenuated. As a result, it is implausible to ignore this concern.

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).

In most applications, estimators people use 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 value (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 σ^2_k. This is more realistic; it is implausible that treatment effect has the same asymptotic variance across subgroups. 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_γ^2 *be* $Var(\gamma_k)$ *and* $\lambda = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \overline{\sigma_{uk}^2}}$*. 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 need to 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 [7]. 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

[7] 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.

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

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

with the asymptotic variance

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

where $\hat{\xi}_k = \frac{(\hat{\gamma}_k - \overline{\hat{\gamma}_k})(\hat{\tau}_k - \hat{\beta}\gamma_k - \hat{\delta}) + \hat{\beta}\sigma_{uk}^2}{\sigma_{uk}^2 - \overline{\sigma_{uk}^2}}$, 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 $g(\sigma_{uk}^2)$. 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, $\sigma_{uk}^2 = 0$. We will show both estimates in the simulation and application and wrap them in the R package.

5.2 Individual Level Estimator

Recall we use $G_k = \{T, X_1, X_2, ..., X_l\}$ 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

6.1 Simulation

In this section, we use Monte Carlo simulation to examine the performance of BCES and SIMEX estimators, especially under the small sample size. We are particularly interested in the small sample size because, in practice, it is not quite easy to detect a large number of subgroups. The data are simulated with the following process:

$$\tau_k = 4 + \beta\gamma_k + N(0,1) \tag{29}$$

$$\sigma_{uk}, \sigma_{vk} \sim Gamma(shape = 1, rate = 1) \tag{30}$$

$$\hat{\gamma}_k = \gamma_k + N(0, \sigma_{uk}^2) \tag{31}$$

$$\hat{\tau}_k = \tau_k + N(0, \sigma_{vk}^2) \tag{32}$$

where true γ_k is drawn from $N(2,1)$.

We first set $\beta = 0$ and compare it with the theoretical rejection rate (i.e.,size) 0.05 (two-tailed test). The left part of the table 1 shows the simulation results. In general, the rejection rate of BCES with asymptotic variance perform worst, reaches 0.422 with sample size 5. BCES with non-parametric pairs bootstrap moderately reduce the over rejection in large sample size. BCES with restricted wild bootstrap and SIMEX performs much better in the samll sample size. In particular, the empirical rejection rate of SIMEX

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.

is almost equal to the theoretical reaction rate (0.05) with 5 observations, the extremely small sample size. In general, this suggests that we should not use original BCES without adjustment.

We then examine the statistical power of estimators. We let $\beta = 2$ so that the average indirect effect is around 50 percent of the total treatment effect, which is a reasonable benchmark. The results are shown in the right part of the table 1. We observe a different trend; BCES with pairs bootstrap has larger power when the sample size is small while SIMEX and restricted wild bootstrap BCES have great power when the sample size is lager than 10. Generally, SIMEX has the best performance in the simulation.

One important question is which sample size, the sample size in each subgroup or the number of sample groups, is critical to the statistical power. This is particularly important in practice, especially in experimental design, which can

We address this question Because of no analytical

use the fact that $\sigma_{uk} \approx \frac{c_k}{\sqrt{n_k}}$.

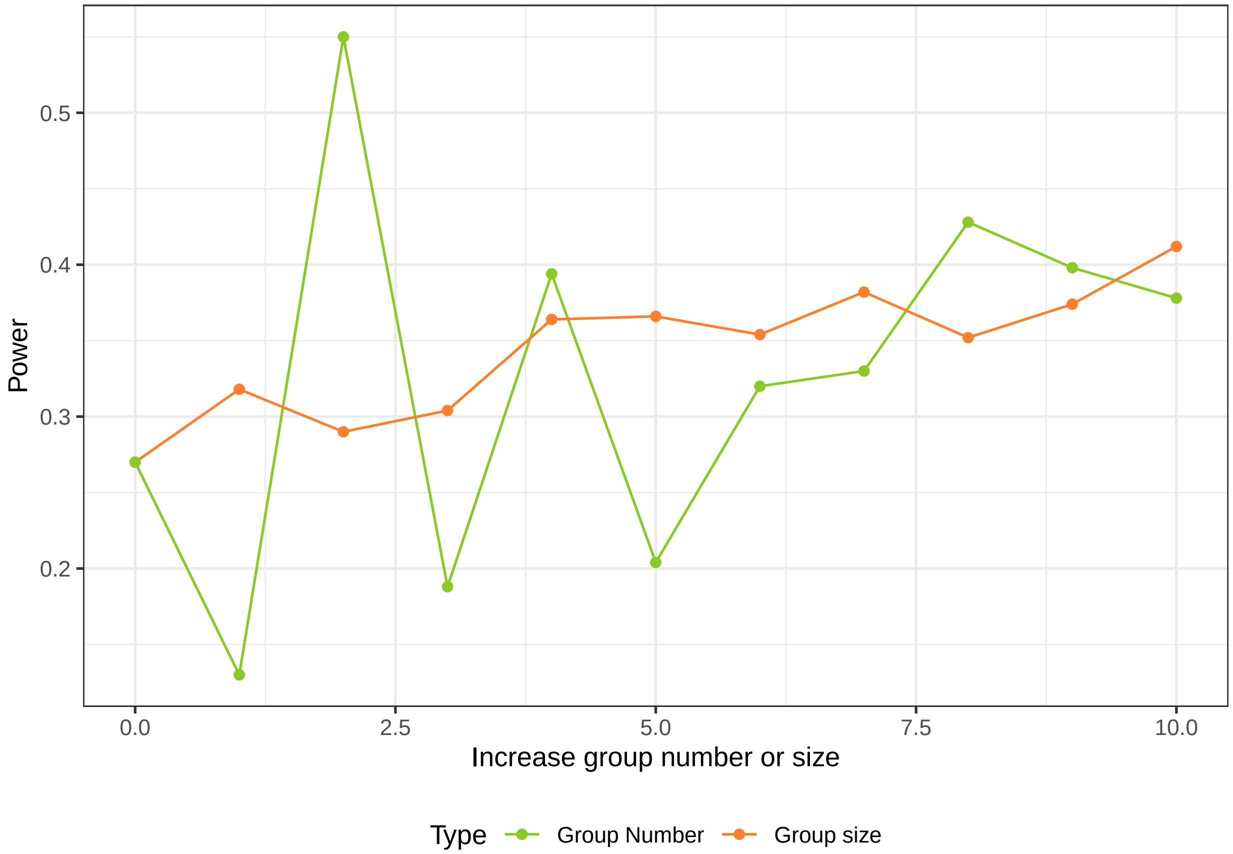

Figure 6: Power analysis: Group number and size

	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: Aggregate-level Data with Experiments from Slough et al. (2021)

Evidence in Governance and Politics (EGAP) [8] 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, they conducted six harmonized experiments with the same 'meta' treatment (community monitoring) but heterogeneous CPRs and treatment sub-types, as shown in the 2.

In their paper, they report effects on multiple outcome variables (including resource use, user satisfaction, user knowledge about community's CPRs, and resource stewardship), and also explore mechanisms: how monitoring affects those outcomes through different channels. However, they simply examine the treatment effects on mediators; this does not necessarily inform the exact causal mediation effects. Instead, we will use their data to exactly quantify the mediation effect. Particularly, we explore two mechanical questions:

(1) how the monitoring effects on resource use are mediated by user scrutiny on CPR management authorities; and

(2) how the monitoring effects on user knowledge about their community's CPRs are

[8] https://egap.org/

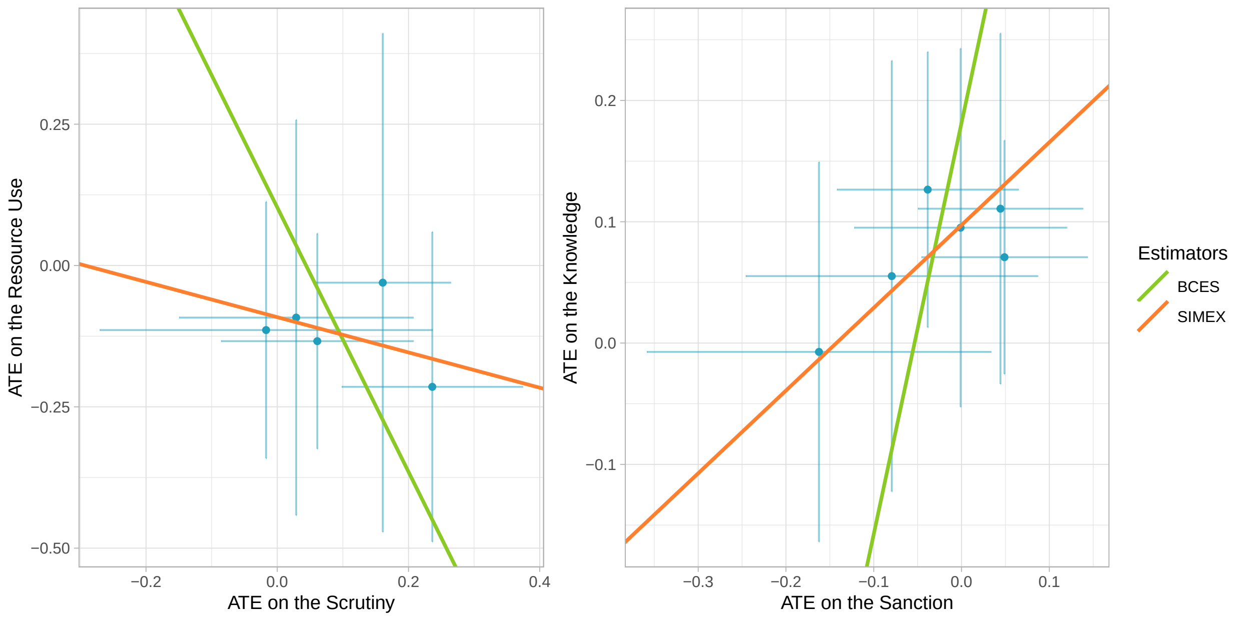

Figure 7: Heterogeneous Subgroup Design

mediated by perceptions of the likelihood of being sanctioned for misuse of CPRs.

The six experiments naturally provide us with six subgroups. In practice, researchers can further define subgroups within each site; however, we will only use those six subgroups to illustrate that, even with such small sample sizes, our estimators can still be applied. We will use both BCES and SIMEX estimators. Correspondingly, we require data on: (1) average treatment effects on both the outcome and the mediator; (2) standard errors. The estimated β_{bces} (pairs), β_{simex} and corresponding data points are shown in figure 7.

On the left panel, blue points with error bars denote average effects on mediator (scrutiny) and outcome (resources use) in five experiments ("in China, the measure of scrutiny was not measured at the citizen level like in the other sites and is thus omitted"). As we can see, in general, those original estimates are not quite precise (due to the moderate sample size in the 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. Therefore, we only draw the green line in the figure. 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. 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, on the right panel, we illustrate the average treatment on the mediator (sanction) and the outcome (knowledge). Even with six observations, two estimates are still significantly congruent on the positive sign ($\beta_{simex} = 0.67, p = 0.05(pairs)$, and $\beta_{bces} = 3.39, p = 0.02$). 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; then we conclude that the indirect effect through the likelihood of being sanctioned can account for 27% of the total monitoring effect on the increasing knowledge about community's CPR.

6.3 Application II: Individual level data

TBD. (Any suggestion?)

7 Extension

1,sensitivity

1, more covariates 2, sensitivity analysis 3, machine learning etc. 4, panel data

DAG cyrus case

5, mention: weight colin p817: weights should be used if regression is viewed as a tool to describe population responsees but need not be used if the regression model is assumed to be the correct structural model.

8 Conclusion

Social science is rooted in two pivotal foundations: theory and empirical research. A concerning observation indicates a growing detachment between these activities. Starting from Merton (1968), the development of middle-range theory to explain and generalize observed empirical phenomena remains a fundamental and predominant objective in the field of social science. Understanding the causal mechanism is indispensable for building a robust theory; moreover, it aids in predicting predict the causal pattern in new environments. Mediation analysis offers powerful statistical tools for quantifying the causal mediation effect. However, a straightforward method that does not require strong sequential ignorability assumptions and can be easily applied in various empirical studies is 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, requiring no advanced techniques beyond simple linear regression.

Although our method does not require onerous ignorability assumptions on the mediator, it is crucial to recognize that it is not the one-size-fits-all solution for mediation analysis. First, it is not entirely model-free under specific situations. For example, we may need to approximate the correlation between β_k and γ_k. However, interpreting this parametric identification as a failure would be misguided; "In modern econometrics, this is generally viewed as a second-best solution as identification has been achieved only through the use of an arbitrary and unverifiable parametric assumption" (Hansen 2022, p56). Second, many areas that need to be further explored still exist. For example, how to extend the method to non-binary treatment? Is it possible to address multiple correlated mechanisms? For the individual-level estimator, can we integrate other causal identification strategies except for IV? Third, our method provides another potential bridge to link causal mediation and causal moderation. We are convinced that many other promising combinations await exploration.

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.

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*.

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.

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.

Merton, Robert King (1968). *Social theory and social structure*. Simon and Schuster.

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.

Supplementary Information

A Multiple Mechanisms

In this section, we consider the general case allowing multiple independent mechanisms. We let M_1 be the mediator of interests, and let M_{-1} be other mediators. We define the overall effect by τ^i:

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

The direct effect is defined as

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

Because we allow multiple mechanisms, apart from the convention, we use $j-$ and $j+$ to denote index $h \in J$ such that $h < j$ and $h > j$, respectively. For the indirect effect, we define as

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

The overall causal effect can be decomposed as:

$$\tau^i = \delta^i(t') + \sum_{j=2}^{J} \eta_j^i(t', t) + \eta_1^i(t', t)$$

To verify it, we let $t' = 1$ and $t = 0$;

$$\begin{aligned}
\tau^i &= \delta^i(1) + \sum_{j=1}^{J} \eta_j^i(1,0) \\
&= Y^i(1, M_1^i(1), ..., M_J^i(1)) - Y^i(0, M_1^i(1), ..., M_J^i(1)) \\
&+ Y^i(0, M_1(1), ..., M_j(1),) - Y^i(0, M_1(0), ..., M_j(1)) \\
&+ Y^i(0, M_1(0), M_2(1) ..., M_j(1),) - Y^i(0, M_1(0), M_2(0), ..., M_j(1)) \\
&+ ... \\
&= Y^i(1, M_1^i(1), ..., M_J^i(1)) - Y^i(0, M_1^i(0), ..., M_J^i(0))
\end{aligned}$$

Basically, the first term in each line is canceled out by the second term in the previous line.

Notably, previous definitions are not general enough. For example, in the direct effect, we require all mediators to take potential outcomes under treatment t'. In general, different mediators can take different potential outcomes. Similarly, for the indirect effect η_j, different mediators other than j can take any possible potential outcomes. But whatever potential outcomes they take, our results hold if the mechanism of interests is additively separable from other mechanisms:

$$\tau^i = (\delta^i + \sum_{j=2}^{J} \eta_j^i) + \eta_1^i \tag{33}$$

The average level decomposition has the similar form: $\tau = (\delta + \sum_{j=2}^{J} \eta_j) + \eta_1$.

B Proof of Proposition 1

Proof. We first decompose the average total causal effect τ as follows:

$$
\begin{aligned}
\tau(t,t') &= \mathbb{E}[Y^i(t,M^i(t)) - Y^i(t',M^i(t))] + \mathbb{E}[Y^i(t',M^i(t)) - Y^i(t',M^i(t'))] \\
&= \mathbb{E}[Y^i(t,M^i(t)) - Y^i(t',M^i(t))] + \frac{\mathbb{E}[Y^i(t',M^i(t)) - Y^i(t',M^i(t'))]}{\mathbb{E}[M^i(t) - M^i(t')]} \times \mathbb{E}[M^i(t) - M^i(t')] \\
&:= \delta + \beta\gamma
\end{aligned}
$$

Then, given the random sample $(\tau_k, \delta_k, \beta_k, \gamma_k)$, we convert it to be simple linear regression

$$
\tau_k = \mathbb{E}\delta_k + \beta_k\gamma_k + \varepsilon_k \tag{34}
$$

where $\varepsilon_k = \delta_k - \mathbb{E}\delta_k$.

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

We first to show result (2) that $\hat{\beta} \to \mathbb{E}\beta_k$. Note that

(1) By construction, $\mathbb{E}\varepsilon_k = \mathbb{E}\delta_k - \mathbb{E}\delta_k = 0$;

(2) Assumption 2 implies that $\mathbb{E}[\gamma_k\epsilon_k] = \mathbb{E}[\gamma_k(\delta_k - \mathbb{E}\delta_k)] = Cov(\gamma_k, \delta_k) = 0$.

$$
\begin{aligned}
\hat{\beta} &= \frac{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)\tau_k}{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} && (35) \\
&= \frac{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)(\overline{\delta}_k + \beta_k\gamma_k + \varepsilon_k)}{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} && (36) \\
&= \frac{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)\beta_k\gamma_k}{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} + \frac{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)\varepsilon_k}{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} && (37) \\
&\xrightarrow{p} \frac{\mathbb{E}[(\gamma_k - \overline{\gamma}_k)\gamma_k\beta_k]}{Var(\gamma_k)} + \frac{\mathbb{E}[(\gamma_k - \overline{\gamma}_k)\varepsilon_k]}{Var(\gamma_k)} && (38) \\
&= \frac{\mathbb{E}[(\gamma_k - \overline{\gamma}_k)\gamma_k]\mathbb{E}\beta_k}{Var(\gamma_k)} + \frac{\mathbb{E}\gamma_k\varepsilon_k}{Var(\gamma_k)} && (39) \\
&= \mathbb{E}\beta_k && (40)
\end{aligned}
$$

where line (37) comes from $\overline{\delta}_k \sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k) = 0$, line (38) is implied by Slutsky's Lemma, (39) is implied by $\mathbb{E}\varepsilon_k$=0 and assumption $\beta_k \perp\!\!\!\perp \gamma_k$, the last line is implied by $\mathbb{E}[\gamma_k \epsilon_k] = 0$,

Result (1) trivially follows the same logic.

□

C Proposition 1 under Multiple Mechanisms

In the main text, when we discuss our novel decomposition and identification assumptions, we consider "no interaction effect" so that the decomposition is unique. Here, we can slightly relax it to be "no interaction effect with respect to M_1".

Assumption C1 (No interaction effect with respect to M_1)**.** *For any $t_j \in \{0,1\}$ where $j = 1,2,..,J$, $\eta_1(t_1, M_1(1), M_2(t_2), ..., M_2(t_j)) - \eta_1(t_1, M_1(0), M_2(t_2), ..., M_2(t_j)) = B$*

In other words, the assumption allows any possible interaction effect among the treatment T and other mediators M_{-1}; however, the effect of M_1 does not depend on them. Under this assumption, without loss of the generality, we use Δ to denote $\delta + \sum_{j=2}^{J} \eta_j$ and thus

$$\tau = \Delta + \eta_1 \tag{41}$$

Similarly, to have a unique form of γ (the treatment effect on the mediator M_1), we also need a kind of "no interaction effect."

Assumption C2 (No interaction effect)**.** *For any $t_j \in \{0,1\}$ where $j = 2,..,J$,*

$$\mathbb{E}Y^i(M_1(1), M_2(t_2), ..., M_2(t_j)) - \mathbb{E}Y^i(M_1(0), M_2(t_2), ..., M_2(t_j)) = D$$

Under the above two "no interaction effect" assumptions, subsequently, we can modify the Proposition 1 as follows:

Proposition 3. *Let (τ, Δ, γ) are random variables. Given the random sample $(\tau_k, \gamma_k)_{k \in K}$. Suppose $Var(\gamma_k) > 0$ and $Cov(\gamma_k, \Delta_k) = 0$.*

Considering the 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 $\hat{\beta}\gamma_k$.

The key difference between the above-modified proposition and the original one is the identification assumption. Here, we need $Cov(\gamma_k, \Delta_k) = 0$. It means that the treatment effect on the mediator of interest is not correlated to the direct effect and other mechanisms.

D Proof of Unbiasedness

Proposition 4. *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 following two assumptions hold:*

(1) (Variance) $Var(\gamma_k) > 0$;

(2) (Mean Independence) $\mathbb{E}[\delta_k | \gamma_k] = \mathbb{E}[\delta_k]$

Considering the 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* $\mathbb{E}\hat{\beta} = \beta$;

(2) If β_k *is a random variable, then* $\mathbb{E}\hat{\beta} = \mathbb{E}\beta_k$ *under assumption* $\mathbb{E}[\beta_k | \gamma_k] = \mathbb{E}\beta_k$,

and thus η_k *is unbiased.*

Proof. For unbiasedness, note that by construction, $\mathbb{E}\varepsilon_k = \mathbb{E}\delta_k - \mathbb{E}\delta_k = 0$ and thus with mean independence assumption (2) we have $\mathbb{E}[\epsilon_k | \gamma_k] = \mathbb{E}\varepsilon_k = 0$. From line (37), we take the expectation given observed $\gamma_1, \gamma_2, ..., \gamma_K$,

$$\mathbb{E}[\hat{\beta} | \gamma_1, \gamma_2, ..., \gamma_K] = \mathbb{E}[\beta_k] \frac{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)\gamma_k}{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)^2} + \frac{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)\mathbb{E}[\epsilon_k | \gamma_k]}{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)^2} \quad (42)$$

$$= \mathbb{E}\beta_k \quad (43)$$

Result (1) trivially follows the same logic.

□

E Proof of Proposition 2

Proof. Firstly, We calculate the expectation of $\hat{\gamma}_k^2 = \gamma_k^2 + 2\gamma_k u_k + u_k^2$. Let $\mu_\gamma = \mathbb{E}\gamma_k$.

For each part, we have

$$\mathbb{E}\gamma_k^2 = \sigma_\gamma^2 + \mu_\gamma^2 \tag{44}$$

$$\mathbb{E}2\gamma_k u_k = 0 \text{ by } Cov(\gamma_k, u_k) = 0 \tag{45}$$

$$\mathbb{E}u_k^2 = \sigma_{uk}^2 \tag{46}$$

Therefore, $\mathbb{E}\hat{\gamma}_k^2 = \sigma_\gamma^2 + \mu_\gamma^2 + \sigma_{uk}^2$.

Now, considering the estimator,

$$\hat{\beta} = \frac{\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k}{\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{47}$$

$$= \frac{\frac{1}{K}\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})[\mathbb{E}\delta + \gamma_k\beta_k + (\varepsilon_k + v_k)]}{\frac{1}{K}\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{48}$$

To see the convergence of the denominator, we re-write it as $\frac{\sum \hat{\gamma}_k^2}{K} - (\frac{\sum \gamma_k}{K})^2$.

Note that $\hat{\gamma}_k^2$ is independent but not identically distributed. When applying Kolmogorov's strong law of large numbers, we need assumption $\sum_{k=1}^\infty \frac{Var(\hat{\gamma}_k^2)}{k^2} < \infty$. Under the assumption, we conclude that

$$\frac{\sum \hat{\gamma}_k^2}{K} \to \sigma_\gamma^2 + \mu_\gamma^2 + \overline{\sigma_{uk}^2}$$

and

$$(\frac{\sum \gamma_k}{K})^2 \to \mu_\gamma^2$$

with continuous mapping theorem. Thus, we have $\frac{1}{K}\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2 \to \sigma_\gamma^2 + \overline{\sigma_{uk}^2}$.

For the numerator, we consider $\frac{\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})\gamma_k}{K}$. Similarly, we find $\frac{\hat{\gamma}_k \gamma_k}{K} \to \sigma_\gamma^2 + \mu_\gamma^2$ and

$\frac{\overline{\hat{\gamma}_k}\gamma_k}{K} \to \mu_\gamma^2$, and thus $\frac{\sum_{k=1}^K(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\gamma_k}{K} \to \sigma_\gamma^2$.

Return to the estimator, we have

$$\hat{\beta} = \frac{\frac{1}{K}\sum_{k=1}^K(\hat{\gamma}_k - \overline{\hat{\gamma}_k})[\mathbb{E}\delta + \gamma_k\beta_k + (\varepsilon_k + v_k)]}{\frac{1}{K}\sum_{k=1}^K(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{49}$$

$$\xrightarrow{p} \lambda\mathbb{E}\beta_k. \tag{50}$$

where we use the same methods in the proof of Proposition 1 and zero covariance $Cov(\gamma_k, v_k) = 0$ in the assumption.

□

F BCES estimator

Ideally, if we have data on the true value (γ_k, τ_k), the OLS estimator is consistent, from Proposition 1 and the proof B:

$$\begin{aligned} \hat{\beta}_{ideal} &= \frac{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)\tau_k}{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)^2} && (51) \\ &\to \frac{\sigma_\gamma^2 \mathbb{E}\beta_k}{\sigma_\gamma^2} && (52) \end{aligned}$$

However, we only observe $(\hat{\gamma}_k, \hat{\tau}_k)$; therefore, the empirical estimator is attenuated, by proof E:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k}{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} && (53) \\ &= \frac{\frac{1}{K}\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})[\mathbb{E}\delta + \gamma_k\beta_k + (\varepsilon_k + v_k)]}{\frac{1}{K}\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} && (54) \\ &\to \frac{\sigma_\gamma^2 \mathbb{E}\beta_k}{\sigma_\gamma^2 + \overline{\sigma_{uk}^2}} && (55) \end{aligned}$$

Therefore, to obtain a consistent estimator, in the denominator, we could subtract $\overline{\sigma_{uk}^2}$. The modified estimator is exactly the BCES estimator:

$$\hat{\beta}_{BCES} = \frac{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k}{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2 - \sum_{k=1}^{K}\sigma_{uk}^2} \tag{56}$$

If we allow correlation between u_k and v_k, we should adjust the numerator as well. Let σ_{uvk}^2 denote the covariance for observation k. The resulting BCES estimator is the same as the one proposed in the Akritas and Bershady 1996:

$$\hat{\beta}_{BCES} = \frac{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k - \sum_{k=1}^{K}\sigma_{uvk}^2}{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2 - \sum_{k=1}^{K}\sigma_{uk}^2} \tag{57}$$

G Details in the Bootstrap

Restricted Wild Bootstrap

- Step 1: Calculate $\hat{\beta}_{BCES}$ and $\sigma_{\hat{\beta}}$ using original sample and form the ratio $t = \frac{\hat{\beta}_{BCES}}{\sigma_{\hat{\beta}}}$.
- Step 2: Impose null hypothesis $\beta_{BCES} = 0$ and calculate associated residuals $(u_1, .., u_k)$
- Step 3: Do B iterations of this step. On the bth iteration:
  - (a) Form the new sample $\{(\hat{y}_1^*, X_1), ..., (\hat{y}_k^*, X_k)\}$ where $\hat{y}_k^* = X_k' \hat{\beta}_{BCES} + u_k^*$. $u_k^* = u_k \lambda_k$. The value of λ_k depends on different method. Rademacher type is $\lambda_k = 1$ with prob 0.5 and $\lambda_k = -1$ with prob 0.5; Six-point type is $\lambda_k = -\sqrt{1.5}, -1, -\sqrt{0.5}, \sqrt{0.5}, 1, \sqrt{1.5}$ with prob $\frac{1}{6}$ respectively.
  - (b) Calculate the ratio $t_k^* = \frac{\hat{\beta}_{BCES}^*}{\sigma_{\hat{\beta}}^*}$ where the numerator and denominator are obtained from the bth pseudo-sample.
- Step 4: Conduct hypothesis testing. Reject null hypothesis at α if $t < t_{[\alpha/2]}^*$ or $t > t_{[1-\alpha/2]}^*$ where the subscript denotes the quantile of $t_1^*, ..., t_k^*$.

Pairs Bootstrap

The same as above except Step 3 (a):

- Step 3: Do B iterations of this step. On the bth iteration:
  - (a) Form the new sample $\{(\hat{y}_1^*, X_1^*), ..., (\hat{y}_k^*, X_k^*)\}$ by re-sampling with replacement K times from the original sample.

H Power analysis

We investigate how the power changes with the number of groups and the number of individuals within each group. In the simulation, we assume that at the beginning, there are 10 groups. Each group has population n. Suppose researchers can enroll another $k * n$ individuals depending on their budget (k is an integer). They face a decision: whether to add k more groups or to increase the group size (i.e. add $\frac{kn}{10}$ individuals to each existing groups).

The data generation process follows the same procedure as outlined in the main text. When we increase the group size, it becomes necessary to enhance the precision of the observed values. To achieve this, we employ the following approximation:

$$se(\hat{\beta}) \approx \frac{c}{\sqrt{n}}$$

Therefore, when adding $\frac{n}{10}$ to each group, the standard error is adjusted to $\frac{\sigma\sqrt{n}}{\sqrt{n+kn/10}}$ where σ is the baseline standard error for γ and τ in the simulation.

In the figure A.1, the vertical axis represents the power. The number on the horizontal line is k, which denotes k more groups (green line) or $kn/10$ more individuals in each group (orange line). The figure does not readily suggest a clear decision regarding the optimal choice between these two options. This ambiguity underscores the importance of conducting a thorough power analysis prior to the experiment to guide such decisions.

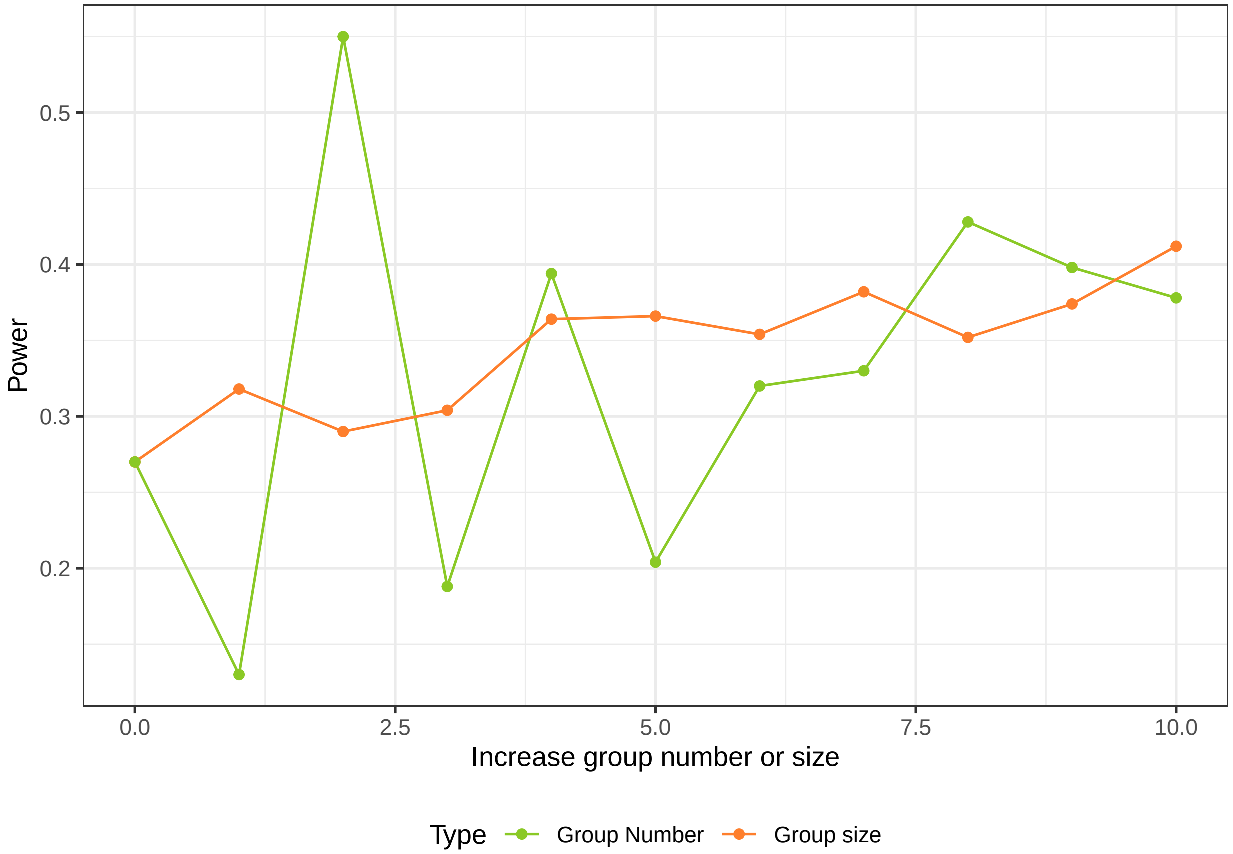

Figure A.1: Power analysis: Group number and size

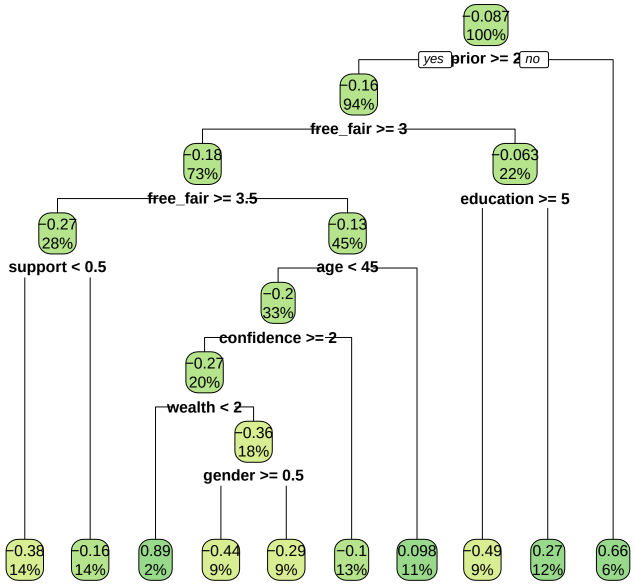

Figure A.2: Heterogeneous Subgroup Design

I Application II

In this section, we employ the full dataset to identify subgroups using causal trees, as depicted in Figure A.2. This process reveals a greater number of subgroups. It's important to note that the number of detected subgroups is influenced by various factors, including the minimum number of observations required for each split. The corresponding estimates are shown in Figure A.3. Upon examination, it is evident that the estimate of β closely aligns with the one discussed in the main text.

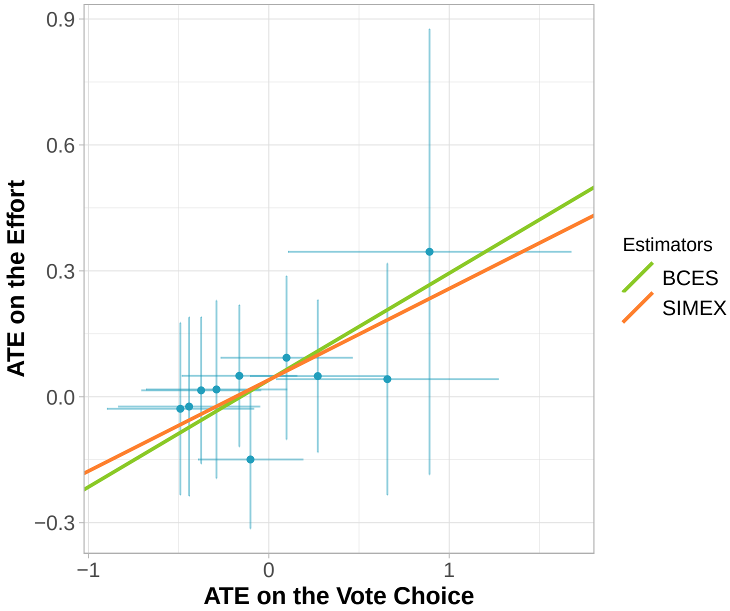

Figure A.3: Heterogeneous Subgroup Design