Extracting Mechanisms from Heterogeneous Effects: An Identification Strategy for Mediation Analysis *

Jiawei Fu†

October 31, 2024

Abstract

Understanding causal mechanisms is crucial for explaining and generalizing empirical phenomena. Causal mediation analysis offers statistical techniques to quantify the mediation effects. However, current methods often require multiple ignorability assumptions or sophisticated research designs. In this paper, we introduce a novel identification strategy that enables the simultaneous identification and estimation of treatment and mediation effects. By combining explicit and implicit mediation analysis, this strategy leverages heterogeneous treatment effects and does not require addressing unobserved confounders. Monte Carlo simulations demonstrate that the method is more accurate and precise across various scenarios. To illustrate the efficiency and efficacy of our method, we apply it to estimate the causal mediation effects in two studies with distinct data structures, focusing on common pool resource governance and voting information. Additionally, we have developed statistical software to facilitate the implementation of our method.

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

*I thank Nathaniel Beck, Adam N. Glymn, Donald Green, Kouske Imai, Dimitri Landa, Connor Jerzak, John Marshall, Cyrus Samii, Yuki Shiraito, Tara Slough, Anna Wilke, Teppei Yamamoto; seminar audiences at Columbia University and New York University; and participants at PolMeth XL and APSA 2023 for helpful feedback.

†Postdoctoral Associate, Yale University jiawei.fu@yale.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 (VanderWeele 2015). Many existing methods rely on the assumption of sequential ignorability, an assumption that may not always be valid, even in completely randomized experiments (Imai et al. 2011). This study proposes an alternative identification strategy to facilitate the causal mediation analysis by exploring heterogeneous treatment effects (HTE).

Mediation analysis has been approached through two main methods. In *explicit mediation analysis*, researchers aim to identify the exact causal mediation effect—that is, the causal effects mediated by a specific proposed mediator. Using either the counterfactual or structural framework, as illustrated in Figure 1, the total treatment effect is decomposed into direct and indirect effects. 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. 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 the second ignorability assumptions is challenging. In the structural modeling

In *implicit mediation analysis*, researchers do not seek to identify the exact causal mediation effect but rather aim to derive qualitative insights into the underlying mechanism. For example, they may estimate the correlation between the treatment and mediator or between the mediator and the outcome (Blackwell et al. 2024). The most popular approach utilizes HTEs. The underlying intuition is that if a specific mechanism is active

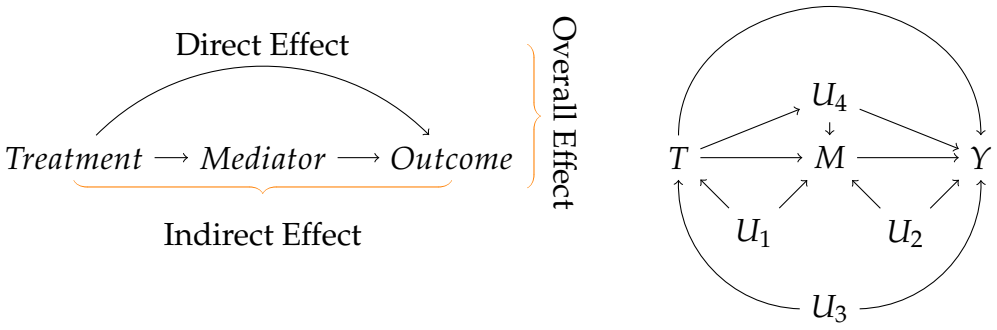

Figure 1: The left panel illustrates the basic decomposition of the overall treatment effect. The right panel is one example of the directed acyclic graph. T is the treatment variable, M is the mediator, Y is the outcome variable, and $U_j, j = 1, 2, 3, 4$ are confounders.

for certain units, it will generate HTEs with respect to a particular covariate. If the mechanism is inert, it will fail to produce HTEs (Fu and Slough 2023). This method is widely used because it does not require measuring the mediator or relying on strong identification assumptions.

Our novel identification strategy incorporates the strength of two approaches and does not require strong exogeneity assumptions. We first decompose the total treatment effect by highlighting the treatment effect on the mediator. As we show in section 3, the new decomposition has the form of simple linear regression, where the 'dependent variable' is the average treatment effect on the outcome and the 'independent variable' is the average total treatment effect on the mediator. To *non-parametrically* identify the causal mediation effects (i.e. the slope), we assume that the average treatment effect on the mediator is not correlated with the error term. In our decomposition, the variance within the error term exclusively comes from other mechanisms. Under this assumption, even if confounders exist between the mediator and the outcome, we can still identify the causal mediation effect.

To implement the strategy, we rely on heterogeneous treatment effects on the mediator as our primary data sources. If the same research is conducted several times in different

locations, we naturally obtain different treatment effects from those studies. In section 4, we propose two more research designs to exploit those HTEs. In the first design, we suggest researchers use pre-treatment covariates to identify those subgroups, probably with 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 variation in treatment intensity. For example, in a Get-Out-The-Vote (GOTV) experiment, researchers might randomly assign different numbers of canvassing mails. By manipulating the magnitude or intensity of the treatment across different arms, we can induce varying treatment effects on the mediator, a design we refer to as the *Multiple Treatment Meta Design*.

Because average treatment effects are estimated from the design, we need to modify the simple OLS estimator to account for the measurement error. Our main estimator builds on the well-known simulation-extrapolation estimator (Cook and Stefanski 1994). Detailed methodologies for estimation and inference are provided in section 5. We then use Monte Carlo simulation to examine the performance of those estimators in section 6. When compared to the identification and estimation methods under sequential ignorability, our method demonstrates greater robustness and efficiency. The efficiency gain observed is analogous to that achieved through (post-)stratification as opposed to completely randomized experiments. 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 group-level 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. In section 7, we offer a range of valuable extensions designed to assist applied researchers in meeting our core identification assumptions through both

the design stage and the data analysis stage.

In general, we introduce an alternative identification strategy for causal mediation analysis that combines explicit and implicit approaches. Instead of relying on unconfoundedness assumptions, our method leverages HTEs and the theoretical structure of mechanisms. No single method is suitable for all study types: identification based on "exogeneity" or "ignorability" should be applied when researchers have sufficient data and understanding of potential confounders. Conversely, when confounders are unclear or unmeasured, our approach—relying on HTEs and knowledge of mechanisms—may be a more effective choice.

2 Causal Mediation Framework

In this section, we review basic definitions and notations for mediation analysis, if necessary, with the help of directed acyclic graphs (DAG).[1]

2.1 Running Example

Throughout this paper, we will use the Get-Out-The-Vote experiments as running examples to illustrate various concepts. Consider the scenario where researchers have observed that treatments such as door-to-door canvassing, phone calls, or mailings can significantly increase voter turnout (Green and Gerber 2019). A critical question arises: through what mechanisms, do these treatments influence turnout? Mediation analysis helps us quantify the average causal effects along these pathways.

To better understand the mechanisms involved in this example, we draw on politi-

[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.

cal economy theory. According to Riker and Ordeshook (1968), there are four elements that influence a voter's decision to abstain or vote, as shown in the Figure 2. (1) Cost of Voting (C): This includes costs related to information gathering, registration, and transportation, among others. Generally, higher costs discourage turnout. (2) Civic Duty (D): In democracies, voting is often portrayed as a civic responsibility. D captures the utility derived from fulfilling this civic duty; fail to fulfill the duty may lead to social pressure as dis-utility. (3) Utility Difference (B): Voters are more likely to turn out if the utility gained from their preferred candidate winning is significantly outweighs the utility loss if they lose. (4) Pivotality (p_i): This refers to the probability that a voter's vote could decisively impact the election outcome, with $p_i \in [0, 1]$ representing the probability. A lower likelihood of affecting the result may decrease the motivation to vote.

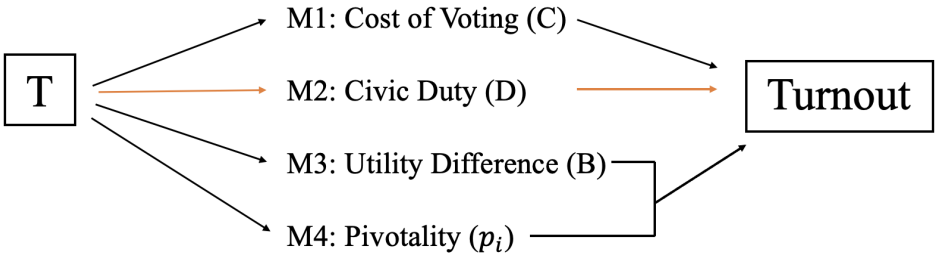

Figure 2: Mechanism for Turnout

Consequently, a voter prefers to vote over abstaining if and only if $p_i B + D - C \geq 0$; that is, the utility from voting outweighs the costs. These four elements—cost, civic duty, utility difference, and pivotality—represent the mechanisms that can influence turnout. As the formula clearly demonstrates, Mechanisms 3 and 4 are correlated. In our causal mediation question, we specifically focus on the civic duty mechanism, measured by social pressure, to quantify its causal effect on voter turnout.

2.2 Causal Decomposition

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. The overall treatment effect of T on Y for individual i, denoted by τ^i, is represented by

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

For the total effects, mediators M should consider potential outcomes under the treatment status. For instance, $M_1^i(1)$ represents the potential social pressure when the treatment $T = 1$.

Following Pearl (2001) and Robins and Greenland (1992), total treatment effects can be decomposed into natural direct and indirect effects (Robins and Greenland 1992). We call $\delta^i(t) = Y^i(1, M^i(t)) - Y^i(0, M^i(t))$ the natural direct effect for $t = 0, 1$, where the mediator is set to the value it would have been under treatment t. It captures the effects that are not transmitted by the mediator of interest. [2]

Similarly, $\eta^i(t) = Y^i(t, M^i(1)) - Y^i(t, M^i(0))$ is the natural indirect effect for $t = 0, 1$. It denotes the treatment effect through the mediating variable. In our example, this refers to how the decision to turn out changes in response to the shift in social pressure from what it would be under the control condition ($M^i(0)$) to the treatment condition ($M^i(1)$), while holding the treatment status constant at t. See more discussions in supplementary materials (SI) A.

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

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

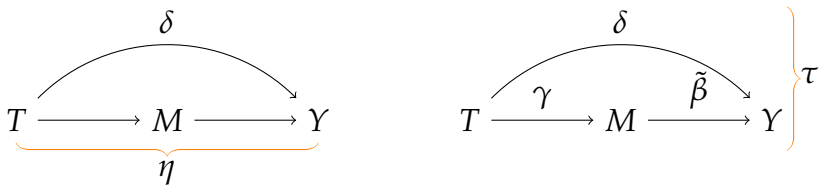

Figure 3: Decompositions with Counterfactual and Structural Approaches

treatment effect: $\tau := \mathbb{E}[\tau^i]$, δ as the average direct effect: $\mathbb{E}[\delta^i]$, and η as the average indirect effect for mechanism represented by the mediator M: $\eta := \mathbb{E}[\eta^i]$. As is standard, illustrated in the left DAG in Figure 3, we can then decompose the total causal effect into the sum of natural direct and indirect effects:

$$\tau = \delta(t) + \eta(1 - t)$$

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:

$$Y = \alpha_1 + \delta T + \tilde{\beta} M + \varepsilon_1 \tag{1}$$

$$M = \alpha_2 + \gamma T + \varepsilon_2 \tag{2}$$

Replacing M (equation (2)) in (1), we obtain the reduced form as follows:

$$Y = (\alpha_1 + \alpha_2\tilde{\beta}) + (\delta + \tilde{\beta}\gamma)T + (\varepsilon_1 + \tilde{\beta}\varepsilon_2) \tag{3}$$

$$:= \alpha_3 + \tau T + \varepsilon_3 \tag{4}$$

Traditionally, parameter before the treatment T, i.e., $\tau = \delta + \tilde{\beta}\gamma$ in (4) 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 3 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. As demonstrated below, the re-introduction of γ into the counterfactual decomposition, generates new insights about the mediation analysis.

2.3 Current Practice

Before demonstrating our new identification strategies, it is crucial to review the current state of the art. In the explicit mediation analysis, we require the sequential ignorability to non-parametrically identify the causal mediation effect. Multiple versions of sequential ignorability assumption exist. 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:

$$\{Y^i(t', m), M^i(t)\} \perp\!\!\!\perp T^i | X^i = x \tag{5}$$

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:

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

In the assumption (6), the mediator takes the value at the "current" treatment assignment t, but the potential outcome is under treatment assignment t'. For example, it requires that the potential social pressure under treatment condition is independent of the potential turnout under control condition, given the individual is under the treatment status and has pre-treatment variables $X^i = x$. The different indices make it hard to understand and 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 [3], 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).

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). 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 (6) and allow post-treatment confounders in X: $Y^i(t, m) \perp\!\!\!\perp M^i | T^i = t, X^i = x$. 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.

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

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. Because of the challenge, in practice, researchers often rely on modeling assumptions to estimate the causal mediation effect. The traditional choice is the linear regression model (equation (1)-(4)). 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.

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:

$$\mathbb{E}[\tilde{\beta}^i \gamma^i] = \mathbb{E}[\tilde{\beta}^i] \times \mathbb{E}[\gamma^i] + Cov(\tilde{\beta}^i, \gamma^i) \tag{7}$$

If the covariance is not zero, then the product estimator is biased. From equation (7), 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 (1): $Y = \alpha_1 + \delta T + \tilde{\beta} M + \theta TM + \varepsilon_1$ (Imai et al. 2010a; Preacher et al. 2007).

Generally, explicit 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 4, 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.

Given these challenges, researchers frequently rely on implicit mediation analysis (Bullock and Green 2021). Although it does not necessarily reveal the causal mediation effect, it can provide useful qualitative insights into the mechanism. However, many practices incorporate unstated assumptions. In practice, measuring the mediator is often difficult. In such cases, we can explore whether the proposed mechanism is active by examining HTEs. The intuition is straightforward: if a mechanism explains the causal relationship, we would expect individuals with higher values of a given covariate to exhibit larger treatment effects, as the covariate is assumed to moderate the effect. However, Fu and Slough (2023) find that valid inference of a mechanism from HTEs requires an exclusion assumption, ruling out the possibility of the covariate moderating through alternative pathways. Moreover, if the outcome variable is binary, even when the exclusion assumption holds, the presence of HTEs does not necessarily indicate mechanism activation. When the mediator is measured, one approach is to test whether the treatment affects the proposed mediator; Blackwell et al. (2024) clarify that a monotonicity assumption is needed to draw sharper conclusions.

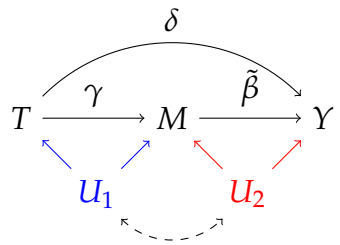

Figure 4: Mediation analysis generally requires addressing confounders U_1 and U_2.

2.4 Other Identification Strategies

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.

By observing Figure 4, 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. [4] 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.

3 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

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

consider the "no interaction effect" situation in the maintext ($\delta = \delta(1) = \delta(0)$ and $\eta = \eta(1) = \eta(0)$). See SI B for the discussion on the interaction effect. The identification strategy begins with a straightforward transformation of this decomposition.

$$\tau = \mathbb{E}[Y^i(1, M^i(0)) - Y^i(0, M^i(0))] + \mathbb{E}[Y^i(1, M^i(1) - Y^i(1, M^i(0))] \tag{8}$$

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

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

$$:= \delta + \beta\gamma \tag{11}$$

The first two lines are two decompositions. In the line (10),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. If researchers plan to use SEM to conduct mediation analysis, under the linear model (1) and (2), we can easily see that $\beta = \tilde{\beta}$. Therefore, β can be interpreted as the effect of the mediator M on the outcome Y [5]. Finally, we use simple notation to represent the final decomposition $\tau = \delta + \beta\gamma$.

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

[5]Under linear SEM, we implicitly assume ε_1 and ε_2 are independent. From equation (1) and (2), we observe

$$\begin{aligned} \mathbb{E}[Y^i(0, M^i(1))] &= \alpha_1 + \tilde{\beta}(\alpha_2 + \gamma) \\ \mathbb{E}[Y^i(0, M^i(0))] &= \alpha_1 + \tilde{\beta}\alpha_2 \end{aligned} \tag{12}$$

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

populations and even among individuals. This implies that the γ is a random variable (We will discuss how to get this sample in the section 4). Therefore, equation (11) can be written as $\tau_k = \delta_k + \beta\gamma_k$. We use subscript k to emphasize that they are random rather than fixed values. Next, we complete the transformation by adding and subtracting the expectation of δ_k.

$$\tau_k = \delta_k + \beta\gamma_k \tag{13}$$

$$= \mathbb{E}[\delta_k] + \beta\gamma_k + (\delta_k - \mathbb{E}[\delta_k]) \tag{14}$$

$$\Rightarrow \tau_k = \mathbb{E}[\delta_k] + \beta\gamma_k + \varepsilon_k \tag{15}$$

In the Line (14), we add and subtract the expectation of δ_k; in line (15), we define $\varepsilon_k = (\delta_k - \mathbb{E}[\delta_k])$. In the structural equation, β, the ratio of the indirect effect to the treatment effect on the mediator is assumed to be constant. In a more general case, β can also be random, and is denoted by β_k.

Equation (15) should be familiar to readers: it is a simple linear regression model. The key assumption for identifying β is that the direct effect δ is uncorrelated with the effect of treatment on mediator γ.

Assumption 1 (Isolated Mechanism)**.** $Cov(\gamma_k, \delta_k) = 0$.

The assumption requires that, γ, the average treatment effect on the mediator of interest, is not correlated with the average direct effect. If there exist multiple mechanisms, generally, we should interpret δ_k as effects from all other possible mechanisms, see SI C and E. Therefore, this assumption implies that the mechanism of interest is somewhat isolated from other mechanisms. This condition is also necessary for implicit mediation analysis using HTEs (Fu and Slough 2023). Intuitively, if the covariate moderates multiple mechanisms, the observed HTEs may not provide information specifically about the mechanism of interest. Similarly, we require that the HTE in γ reveals informa-

tion solely about γ, rather than about other mechanisms. Technically, this 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 $Cov(\gamma_k, \varepsilon_k) = Cov(\gamma_k, \delta_k - \mathbb{E}\delta_k) = Cov(\gamma_k, \delta_k) = 0$.

The validity of this assumption depends on the theoretical framework in question. To clarify, let us revisit our ongoing example. As we introduced earlier, there are four mechanisms—cost, civic duty, utility difference, and pivotality—that influence the decision to vote. If a researcher aims to examine the impact of mailings that encourage voters with the message "DO YOUR CIVIC DUTY—VOTE!" it is plausible to assume that such an intervention mainly affects the civic duty utility. Given that other mediators are distinct and capture different mechanisms, we can reasonably assume that the effect of the canvassing treatment on civic duty is uncorrelated with other mechanisms. However, in general, assumption 1 equires the belief that no other variable concurrently moderates both the mechanism of interest and other mechanisms. We can, however, allow for correlations among other mechanisms, such as the correlation between utility difference and pivotality, as illustrated in 2. In the section 7, we will provide more discussions and techniques on meeting the identification assumption in application.

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 (10) and (11). Suppose we have 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. □

The estimator $\hat{\beta}$ is exactly the simple OLS estimator for the slope. The assumption $Var(\gamma_k) > 0$ is technical; it guarantees that we have "random" observations of γ. The proposition indicates that if we assume the treatment effect on the mediator of interest is not correlated with the effects of other mechanisms, then β can be consistently estimated. Combined with the information of γ, the overall average mediation effect $\mathbb{E}[\eta_k]$ can be estimated by $\hat{\beta}\sum_{k=1}^{K}\gamma_k\mathbb{P}(\gamma_k)$, where $\mathbb{P}(\gamma_k)$ can be consistently estimated by the proportion of sample size k relative to the total sample size. The R package will return all essential statistics. It is important to note that the estimates produced by other mediation analysis techniques, which do not account for heterogeneity, can be interpreted as average effects over implicit heterogeneous effects.

Remark 1. *Whether β is a constant can actually be tested, and we will demonstrate this in the application.*

Remark 2 ($\beta_k \perp \gamma_k$)**.** *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 quadratic linear model $\beta = \theta_0 + \theta_1\gamma + \theta_2\gamma^2$. Then, replace it in the model (15) 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 sta-

tistical 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 F.

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 4). As mentioned in the section 2.3, current methods cannot efficiently address this unconfoundedness problem without further assumptions.

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.

4 How to Obtain Heterogeneous Effects?

To implement the strategy, we need a random sample of treatment effects, (γ_K, τ_K). It is common for the same treatment to generate varying (average) treatment effects across different populations, such as those in different countries or areas. This is similar to getting multiple effects when conducting meta-analysis. Rigorously, we assume there are k independent trials that generate k study-level effects (γ_k, τ_k), where γ_k is assumed

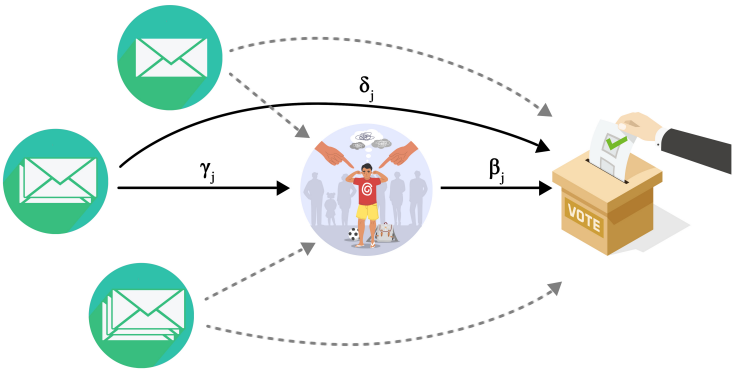

Figure 5: Multiple Treatment Meta Design

to come from a super-population (normal) distribution. In the random-effects model, $\gamma_k = \mu + \Delta_k + e_k$, where μ is the overall effect, Δ_k is the deviation of $k's$-effect from the overall effect, and e_k is the sampling error. As is standard in meta-analysis, we require that Δ_k be independent of study k, or exchangeable in the language of Bayesian Statistics (Higgins et al. 2009). This assumption is critical for statistical inference; therefore, we will proceed under the premise that it holds in the subsequent analysis.

Under the above model, practically, how can we find study-level effects (γ_k, τ_k)? If the same research is conducted several times in different locations, we naturally obtain data from those studies. If not, it is still possible to 'create' multiple studies within a single study. In this section, we introduce two general research designs.

Multiple Treatment Meta Design. The key idea is that we can adjust the treatment to induce heterogeneous effects. For example, in the previous GOTV experiments, researchers could modify the treatment by having some groups receive one mail, while other groups receive two or three mails. See Figure 5. Given the same treatment, varying the intensity or magnitude of the treatment can likely induce different effects. 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 intensity T_j.

Heterogeneous Subgroup Design. This design exploits the heterogeneous effects

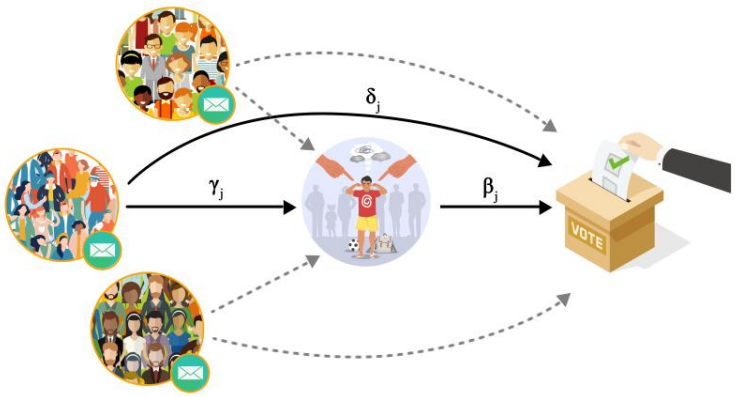

Figure 6: Heterogeneous Subgroup Design

from different population. 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 mailing affects turnout through social pressure in the Get-Out-The-Vote (GOTV) experiment (Gerber et al. 2008), as shown in the Figure 6. 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 should we identify these groups? If several similar studies are conducted in geographically different areas, similar to meta-analysis, then those studies automatically generate a sample of effects. Alternatively, a data-driven method such as causal tree or forest can be used (Wager and Athey 2018) [6].

Synthetic Methods. People may incorporate these two designs and define finer sub-

[6]Technically, the data obtained from such a design represent average conditional effects. The analysis remains valid under the assumption on β, as we take the average across subgroups.

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

5 Estimation and Inference

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.

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)**.**

$$\hat{\gamma}_k = \gamma_k + u_k \tag{16}$$

$$\hat{\tau}_k = \tau_k + v_k \tag{17}$$

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 illustrates the attenuation under our mediation analysis framework and proposes an adjusted estimator.

Proposition 2. *Suppose* (τ_k, γ_k) *satisfies the decomposition* (15)*:* $\tau_k = \mathbb{E}[\delta_k] + \beta_k \gamma_k + \varepsilon_k$*, 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* $\lambda = \frac{\sigma^2_\gamma}{\sigma^2_\gamma + \overline{\sigma^2_{uk}}}$*. 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* $\sum_{k=1}^\infty \frac{Var(\hat{\gamma}^2_k)}{k^2} < \infty$ *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 $\sum_{k=1}^\infty \frac{Var(\hat{\gamma}^2_k)}{k^2} < \infty$. 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 $\lambda = \frac{\sigma^2_\gamma}{\sigma^2_\gamma + \overline{\sigma^2_{uk}}}$, the numerator is the variance of true γ_k; in the denominator, $\overline{\sigma^2_{uk}}$ 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^2_{uk} and therefore can calculate the sample average $\overline{\sigma^2_{uk}}$. However, we need an estimate of the unknown σ^2_γ, 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

[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.

2005).

Although the previous estimator provides insight into the attenuation, practically, we do not recommend directly using it (see also Bowden et al. 2016). Instead, we recommend the simulation-extrapolation estimator (SIMEX) (Cook and Stefanski 1994) [8]. SIMEX 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 approximates the function g and extrapolate back to the case of no measurement error, $\sigma^2_{uk} = 0$.

We employ the intersection-union test for the overall average indirect effect $\mathbb{E}[\eta_k] = \mathbb{E}[\beta_k] \sum_{k=1}^{K} \gamma_k \mathbb{P}[\gamma_k]$, which involves the product of two parameters. For simplicity, we use λ_k to simplify the notation $\mathbb{P}[\gamma_k]$, the proportion of γ_k in the population, and $\hat{\lambda}_k$ as its consistent estimate. The following proposition summarizes how to test the null effect of the overall average indirect effect based on p-values. Let $\hat{\beta}_{SIMEX}$ be the consistent SIMEX estimator of β, and $\hat{\sigma}_{SIMEX}$ be the corresponding standard error. Φ_z is the cumulative distribution function of the standard normal distribution.

Proposition 3. *Assuming the conditions in the proposition of 2 hold, let* $p_\beta = 2 * \Phi_Z(-|\frac{\hat{\beta}_{SIMEX}}{\hat{\sigma}_{SIMEX}}|)$ *and* $p_\gamma = 2 * \Phi_Z(-|\frac{\sum_{k=1}^{K} \gamma_k \hat{\lambda}_k}{\sum_{k=1}^{K} \hat{\lambda}^2 \sigma^2_{vk}}|)$.

Consider the null hypothesis that the overall average indirect effect is zero, $H_0 : \mathbb{E}[\eta_k] = 0$. *Then, rejecting the null hypothesis if* $p_\beta \leq \alpha$ *and* $p_\gamma \leq \alpha$ *constitutes an asymptotic level* α *test.*

In fact, p_β and p_γ are p-values for two sub-parts in the intersection-union test. From the above proposition 3, it follows that the overall p-value associated with the null effect of the overall indirect effect is defined as $\max[p_\beta, p_\gamma]$. The next proposition provides an conservative confidence interval based on the intersection-union test.

Proposition 4. *Assuming the same conditions in the proposition of 3 hold. Define*

[8] We have also considered another estimator in the SI H.

$$a_1 = \hat{\lambda}_0 - \Phi_Z(\frac{1+\sqrt{1-\alpha}}{2}) * \sum_{k=1}^{K} \hat{\lambda}^2 \sigma^2_{vk}$$
$$a_2 = \hat{\lambda}_0 + \Phi_Z(\frac{1+\sqrt{1-\alpha}}{2}) * \sum_{k=1}^{K} \hat{\lambda}^2 \sigma^2_{vk}$$
$$a_3 = \hat{\beta}_{SIMEX} - \Phi_Z(\frac{1+\sqrt{1-\alpha}}{2}) * \hat{\sigma}_{SIMEX}$$
$$a_4 = \hat{\beta}_{SIMEX} + \Phi_Z(\frac{1+\sqrt{1-\alpha}}{2}) * \hat{\sigma}_{SIMEX},$$

$\overline{a} = \max[a_1, a_2, a_3, a_4]$, *and* $\underline{a} = \min[a_1, a_2, a_3, a_4]$.

The at least $(1-\alpha)\%$ *confidence interval for overall average indirect effect is given by* $[\underline{a}, \overline{a}]$, *ensuring* $\mathbb{P}[\underline{a} \leq \mathbb{E}[\eta_k] \leq \overline{a}] \geq 1-\alpha$, *asymptotically.*

Although it is a conservative confidence (in the sense of at least $(1-\alpha)\%$), as we shown in the simulation 6.1, the interval is still smaller than the traditional method under sequential ignorability assumption. Based on both propositions, conducting sub-group inference is also straightforward. We encapsulate estimation and inference in the R package [9]. In the package, we also include Cochran's Q and Higgins & Thompson's I^2 from meta-analysis to assess whether γ_k values exhibit true heterogeneity rather than random error.

6 Simulation and Application

In this section, we employ Monte Carlo simulations to evaluate the effectiveness of our methods by comparing them with the current methods under sequential ignorability. 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.

[9] https://github.com/Jiawei-Fu/mechte

6.1 Simulation

We generate heterogeneous treatment effects for each individual i by assuming 10 subgroups using following simple model:

$$M_{gi} = 1 + \gamma_g * T_i + \kappa * u_i + \epsilon_i^M$$
$$Y_{gi} = 1 + T_i + M_{gi} + \kappa * u_i + \epsilon_i^Y$$

Treatment T_i is randomly drawn from a standard normal distribution, along with two error terms ϵ. The average treatment effect on the mediator M_i varies across groups, with $\gamma_g \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. We also consider an 'unobserved' confounder u_i that simultaneously affects the mediator and outcome, with magnitude $\kappa \in \{0, 1, 2, 3, 4, 5\}$. In Table 1, we show the point estimates and 95% confidence intervals for both methods for different values of κ.

κ	New Method with HTEs (SIMEX)			Traditional Method with Sequential Ignorability		
	ACME	CI low	CI up	ACME	CI low	CI up
0	5.53	5.39	5.66	5.66	5.31	5.96
1	5.63	5.37	5.90	6.18	5.84	6.65
2	5.77	5.31	6.24	7.42	6.88	7.96
3	5.99	5.38	6.63	8.46	7.97	9.09
4	6.08	5.28	6.94	9.18	8.54	9.74
5	6.20	5.14	7.34	9.74	8.89	10.55

Table 1: Comparison between Different Methods

It is evident that when $\kappa = 0$, the sequential ignorability assumption holds. In the first row, therefore, both methods yield estimates that are quite close to the theoretical AMCE of 5.5. However, our method provides tighter confidence intervals. As κ increases, violating the sequential ignorability assumption, our method remains robust to this violation in that the confidence interval still covers 5.5.

In conclusion, our method is robust against the violation of sequential ignorability

and is more efficient. The efficiency gain is intuitive because we explicitly take advantage of the heterogeneous effects. The efficiency gain observed is analogous to that achieved through (post-)stratification as opposed to completely randomized experiments, as discussed in Ding (2023, chapter 5).

Next, we explore how the number of groups and the group size influence statistical power using a similar DGP as before. We increase the total number of individuals from 100 to 200. The additional individuals either form a new group or are distributed among the existing 10 groups to increase group size. In the simulation, we let κ be a random number drawn from a Gamma distribution with a mean of 0.5. Consequently, ACME, which is 0.5, is only half of the direct effect, which is 1. The results are shown in Figure 7. Generally, both increasing the number of groups and the group size can enhance statistical power. However, in this simulation, increasing the number of groups has a more pronounced effect than increasing group size. This suggests that heterogeneity provides more information than increased precision of the estimates. In practice, it is advisable for researchers to conduct similar power analyses to optimize their research designs.

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

Evidence in Governance and Politics (EGAP) [10] 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

[10] https://egap.org/

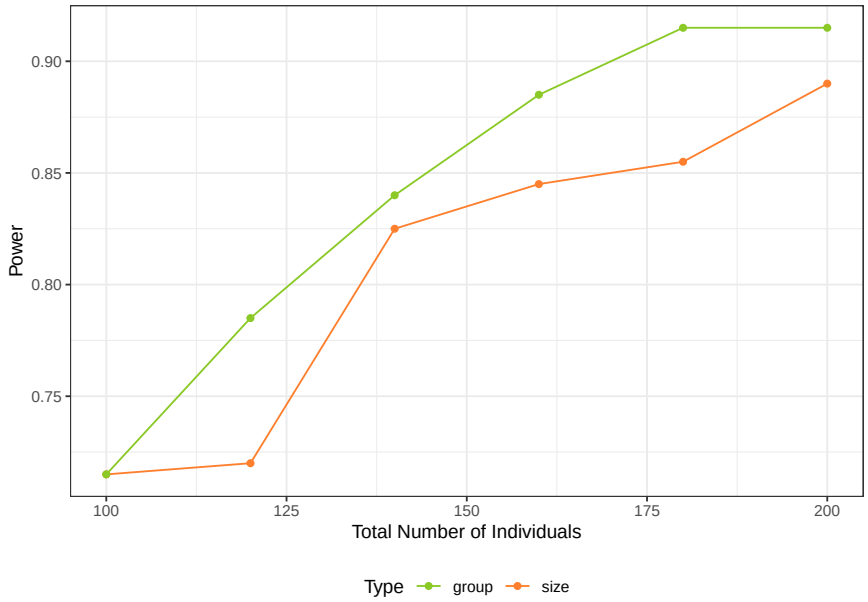

Figure 7: Power Analysis

	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.

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 causal mediation effects. We intend to use their data to illustrate how to apply our causal mediation analysis with aggregate-level data; to be specific, we ask "How does the treatment (i.e. 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 is feasible for researchers to further segment subgroups within each experimental site, our focus will be on these six primary subgroups. To estimate ACME, 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. These data points are represented in the two-dimensional Figure 8, where each dot represents an estimate and each line indicates the 95% confidence interval (CI). Generally, most confidence intervals cross zero, particularly for the treatment effects on the mediator (sanction), as illustrated in Figure 9. Here, no single estimate significantly deviates from zero at the 0.05 level.

However, the data reveal a clear pattern: an increased treatment effect on the perception of sanctions correlates with an increased effect on knowledge about the resource. To quantify the mediation effect, we estimate β, with the estimated slope being $\beta = 0.68$ and the p-value =0.049. This estimate is also depicted by the green line in Figure 8. The figure is also useful for determining whether β is constant. If it is indeed constant, we should expect the data points to display a relatively linear relationship. As previously mentioned, in the structural approach to mediation analysis, we can generally interpret β as the effect of mediator (sanction) on the outcome (knowledge). This confirms our expectation that an increased perception of sanctions enhances the incentive to gain more

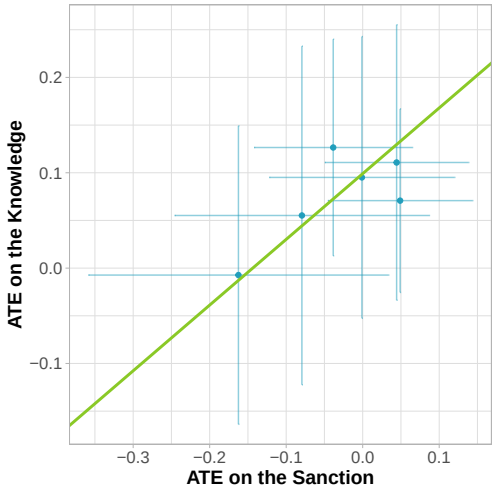

Figure 8: Average treatment effects on the mediator and outcome with standard errors in Six experiments.

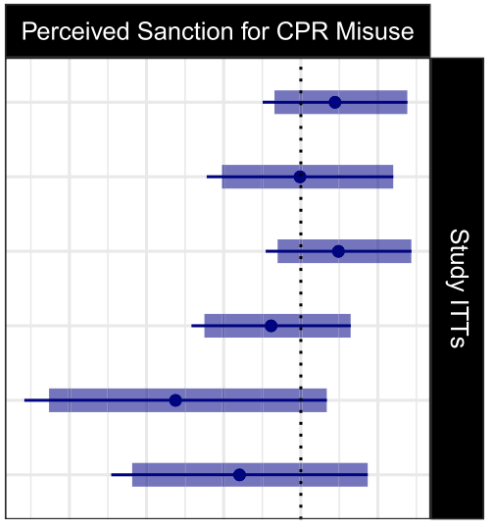

Figure 9: The original figure from Slough et al. (2021) on the average treatment effects on the mediator.

knowledge about common-pool resources.

To obtain the average causal mediation effect, we need to multiply β by γ (the ATE on the mediator) and possibly weight this by the sample size in each experiment. We find that the estimated ACME is 0.01. Since all γ in six experiments are not statistically significant at 0.05 level, it is also challenging to achieve a significant ACME. However, the positive ACME does suggest that the monitoring enhances knowledge about CPRs can be explained by a mediation effect through the perception for CPR misuse.

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 po-

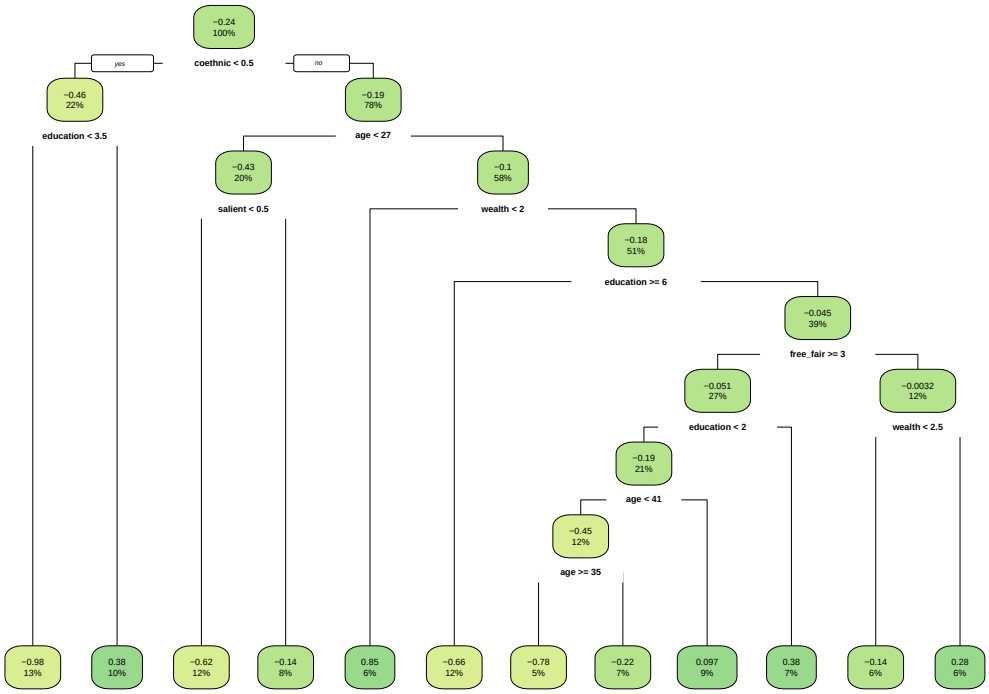

Figure 10: Subgroups detected by Causal trees

tential heterogeneous subgroups. This identification can be achieved 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 10, the informational effect on the perception of effort varies according to factors like coethnic, education, age, wealth status etc. [12]

Next, we estimate the average treatment effect on vote choice across various subgroups, focusing on the "bad information" arm, where the information reveals poor performance by the incumbent. These estimates are then used to calculate the indirect effect using SIMEX, treating them as aggregate-level data. Before that, it is helpful to plot the

[11] For details, please refer to the replication materials.
[12] See more discussions in SI K.

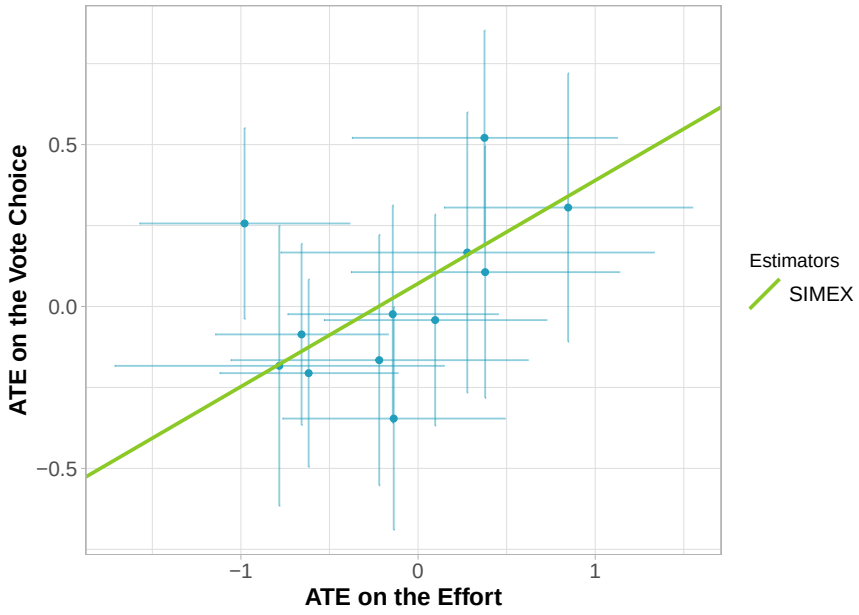

Figure 11: Average treatment effects on the mediator and outcome across subgroups.

data and check for a clear linear relationship. If the data shows linearity, it provides evidence that β is constant, and we do not need to account for non-linearity. The final results are illustrated in Figure 11. We found a clear linear trend and that the estimated β is 0.3, significant at the 0.1 level according to the SIMEX analysis. This suggests that the perception of high effort by the incumbent increases potential votes. The average causal mediation effect is -0.08, also significant at the 0.1 level. Consequently, we deduce that while the overall effect of information may not be substantial enough to detect in the field experiment, there is a significant indirect effect of bad information through voters' perceptions of politicians' effort. Specifically, bad information leads to fewer votes for the incumbent due to perceived lower effort. However, no significant results were detected for the "good information" arm, although the average mediation effect is positively aligned with our expectations.

7 Discussion and Extension

Our identification strategy hinges on a pivotal assumption: $Cov(\gamma_k, \delta_k) = 0$, implying that the treatment effect on the mediator is not correlated with other mechanisms. How can we assess this assumption in practice? The new identification strategy provides two opportunities to address the identification problem—during the design stage and the data analysis stage.

First, researchers can assess this assumption during the design stage, before any experiments or data collection. (1) This can be achieved through a theoretical examination.Understanding the underlying mechanisms necessitates a theoretical foundation. For instance, considering the voter turnout example in section 3, theory identifies four major mechanisms that may influence a voter's decision to abstain or vote: relative payoff from the favored candidate, civic duty, the probability of being pivotal, and voting cost. If theory suggests these mechanisms are uncorrelated, then our assumption is valid. In the turnout example, most theories propose that these mechanisms represent *distinct* aspects of voting calculus and lack evident correlation. For example, the probability of being pivotal is determined by the size of the population, whereas civic duty encompasses all moral considerations, which are unlikely to influence preferences for the candidate. Given a clear treatment, if the mediators capture distinct mechanisms, the identification assumptions are more likely to hold. (2) If conducting an experiment, researchers can control the treatment to minimize potential correlations between γ and δ. For instance, when varying treatment intensity, it is advisable to limit the treatment elements to prevent correlations with other mechanisms.

Second, in the data analysis phase, we can further reduce correlation by carefully selecting covariates to identify heterogeneous treatment groups. Specifically, we select covariates that only moderate the treatment effects on the mediator. For instance, in a

GOTV study, we might hypothesize that door-to-door canvassing could influence both civic duty and perceptions of pivotality. This could occur if the treatment contains multiple elements, such as social pressure from the canvasser combined with information about the polls. It is relatively easy to identify covariates that affect civic duty but not pivotality—for example, neighborhood friendship, education level, or place attachment. These factors likely influence the threshold for civic duty and thus moderates the treatment effect on civic duty due to differing thresholds. However, it is less plausible that neighborhood friendship would moderate perceptions of pivotality. In the era of big data, we typically have access to a large set of covariates, making it easier to find suitable ones for this purpose. A useful approach is to select covariates based on their relationship with the mediator, if mediators are quite distinct, as demonstrated in the example.

Third, a more general solution involves leveraging our novel decomposition, which takes the form $\tau_k = \mathbb{E}[\delta_k] + \beta\gamma_k + \epsilon_k$. This formulation highlights a close analogy to the traditional linear regression model. Should the assumption fail, variables denoted by X, moderating both the treatment effects on the mediator and other mechanisms, may exist. Thus, akin to the approach in traditional linear regression, we *control* for X to 'purify' the error term. As our analysis is based on aggregate-level data (i.e., expected treatment effects), we need account for $\mathbb{E}[X_k]$, the average of X_k in the respective groups, in the linear regression model.

For instance, in our second application detailed in section 6, where we examine how information influences voting choice through voters' perceptions of politicians' efforts, machine learning techniques are utilized to identify heterogeneous subgroups based on pre-treatment covariates. For the sake of argument, let us presume that, beyond perceptions of hard work, the information treatment could also alter voters' beliefs about the competence of incumbent legislators. At least, the belief in incumbent competence is a central mechanism in agency theory (Persson and Tabellini 2002). It may be of con-

cern that the covariate 'partisan' could modify voters' posteriors on both diligence and competence. Therefore, for the seven heterogeneous groups identified, we compute the sample average of 'partisan' in each group, denoted by $\mathbb{E}[X_k]$, and run the linear regression model: $\tau_k = \mathbb{E}[\delta_k] + \beta\gamma_k + \beta_1\mathbb{E}[X_k] + \tilde{\epsilon}_k$. We remain focused on the estimate of β, which is 0.18 with a standard error of 0.06 and a p-value of 0.04. This result closely aligns with the original estimate of 0.17, as expected, since the information treatment seems not to convey information about other mechanisms in this application.

Owing to the challenge of obtaining a large number of observations of average treatment effects (γ_k, δ_k) in practice, indiscriminately adding all covariates in the linear regression model without careful consideration is not advised. This underscores the crucial role of theory in mediation analysis. The decision to include a covariate hinges on theoretical considerations regarding its potential to modify other mechanistic pathways. With theory, researchers can readily identify the most 'important' omitted moderators, as measured by R^2, and incorporate them into the regression to enhance the robustness of their estimates (Oster 2019). With the extension to linear regression, researchers may also apply traditional *sensitivity analysis* to evaluate the robustness of their estimates against unobserved confounders (e.g. Cinelli and Hazlett 2020).

8 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 an alternative identification assumption and strategy that can enable researchers easily estimate causal mediation effects. Within the potential out-

come 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 isolated mechanism assumption, once researchers identify the treatment effects on the mediator and the outcome, our approach can consistently estimate the indirect effect. The proposed strategy enables researchers to address the identification problem in both the design and analysis stages.

While our method reduces the need for unconfoundedness assumptions, it is important to recognize that researchers should select appropriate mediation analysis techniques based on the specific design of their studies. All methods come with their own set of assumptions. Furthermore, 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*.

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". *Proceedings of the 19th international joint conference on Artificial intelligence*, pp. 357–363.

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.

Berger, Roger L and George Casella (2001). *Statistical inference*. Duxbury.

Blackwell, Matthew et al. (2024). "Assumption Smuggling in Intermediate Outcome Tests of Causal Mechanisms". *arXiv preprint arXiv:2407.07072*.

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.

Bullock, John G and Donald P Green (2021). "The failings of conventional mediation analysis and a design-based alternative". *Advances in Methods and Practices in Psychological Science* 4.4, p. 25152459211047227.

Carroll, Raymond J et al. (1996). "Asymptotics for the SIMEX estimator in nonlinear measurement error models". *Journal of the American Statistical Association* 91.433, pp. 242–250.

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

Cinelli, Carlos and Chad Hazlett (2020). "Making sense of sensitivity: Extending omitted variable bias". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.1, pp. 39–67.

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.

Ding, Peng (2023). "A first course in causal inference". *arXiv preprint arXiv:2305.18793*.

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.

Green, Donald P and Alan S Gerber (2019). *Get out the vote: How to increase voter turnout*. Brookings Institution Press.

Higgins, Julian PT et al. (2009). "A re-evaluation of random-effects meta-analysis". *Journal of the Royal Statistical Society Series A: Statistics in Society* 172.1, pp. 137–159.

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". *Statistical Sicence* 25.1, pp. 51–71.

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.

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

Oster, Emily (2019). "Unobservable selection and coefficient stability: Theory and evidence". *Journal of Business & Economic Statistics* 37.2, pp. 187–204.

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.

Persson, Torsten and Guido Tabellini (2002). *Political economics: explaining economic policy*. MIT press.

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

Riker, William H and Peter C Ordeshook (1968). "A Theory of the Calculus of Voting". *American political science review* 62.1, pp. 25–42.

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. 000010220215574679 1329.

— (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 More on Decomposition 2

B Interaction Effect 3

C Multiple Mechanisms 6

D Proof of Proposition 1 8

E Proposition 1 under Multiple Mechanisms 10

F Proof of Unbiasedness 12

G Proof of Proposition 2 13

H BCES estimator 15

I Proof of Proposition 3 16

J Proof of Proposition 4 16

K More on Simulation and Power analysis 18

K.1 Simulation 18

K.2 Power Analysis 19

K.3 Application II 21

A More on Decomposition

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.

Together, we obtain two different decompositions:

$$\begin{aligned} \tau &= \delta(0) + \eta(1) \\ \tau &= \delta(1) + \eta(0) \end{aligned} \tag{18}$$

Assuming no interaction effect exists between the treatment and mediator, the pure 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: $\tau = \delta + \eta$.

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.

B 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:

$$\mathbb{E}[Y] = \alpha_1 + \delta T + \tilde{\beta} M + \theta T M \tag{19}$$

$$\mathbb{E}[M] = \alpha_2 + \gamma T \tag{20}$$

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:

$$\begin{aligned}\tau_1 &= \delta(1) + \eta(0) \\ &= \mathbb{E}[Y^i(1, M^i(1)) - Y^i(0, M^i(1))] + \mathbb{E}[Y^i(0, M(1)) - Y^i(0, M(0))] \\ &= [\delta + \theta(\alpha_2 + \gamma)] + (\tilde{\beta}\gamma)\end{aligned} \tag{21}$$

and

$$\begin{aligned}\tau_2 &= \delta(0) + \eta(1) \\ &= \mathbb{E}[Y^i(1, M^i(0)) - Y^i(0, M^i(0))] + \mathbb{E}[Y^i(1, M(1)) - Y^i(1, M(0))] \\ &= (\delta + \theta\alpha_2) + [(\tilde{\beta} + \theta)\gamma]\end{aligned} \tag{22}$$

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 (15), in the decomposition (21), the "parameter" β is equal to $\tilde{\beta}$. We note that the total direct effect $\delta(1)$ contains γ. However, in the decomposition (22), 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 (22). However, in general, we cannot consistently estimate $\tilde{\beta}$ and thus pure indirect effect $\eta(0)$ with (21) 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).

C 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:

$$\tau^i = (\delta^i + \sum_{j=2}^{J} \eta_j^i) + \eta_1^i \tag{23}$$

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

D 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

$$\tau_k = \mathbb{E}\delta_k + \beta_k\gamma_k + \varepsilon_k \tag{24}$$

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

$$\hat{\beta} = \frac{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)\tau_k}{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} \tag{25}$$

$$= \frac{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)(\overline{\delta}_k + \beta_k\gamma_k + \varepsilon_k)}{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} \tag{26}$$

$$= \frac{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)\beta_k\gamma_k}{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} + \frac{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)\varepsilon_k}{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} \tag{27}$$

$$\xrightarrow{p} \frac{\mathbb{E}[(\gamma_k - \overline{\gamma}_k)\gamma_k\beta_k]}{Var(\gamma_k)} + \frac{\mathbb{E}[(\gamma_k - \overline{\gamma}_k)\varepsilon_k]}{Var(\gamma_k)} \tag{28}$$

$$= \frac{\mathbb{E}[(\gamma_k - \overline{\gamma}_k)\gamma_k]\mathbb{E}\beta_k}{Var(\gamma_k)} + \frac{\mathbb{E}\gamma_k\varepsilon_k}{Var(\gamma_k)} \tag{29}$$

$$= \mathbb{E}\beta_k \tag{30}$$

where line (27) comes from $\bar{\delta}_k \sum_{k=1}^{K} (\gamma_k - \bar{\gamma}_k) = 0$, line (28) is implied by Slutsky's Lemma, (29) 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.

□

E 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 E1 (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

$$\tau = \Delta + \eta_1 \tag{31}$$

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 E2 (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 5. *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.

F Proof of Unbiasedness

Proposition 6. *Let* (τ, δ, γ) *be random variables and as defined in* (10) *and* (11). *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 (27), we take the expectation given observed $\gamma_1, \gamma_2, ..., \gamma_K$,

$$\mathbb{E}[\hat{\beta}|\gamma_1, \gamma_2, ..., \gamma_K] = \mathbb{E}[\beta_k] \frac{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)\gamma_k}{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)^2} + \frac{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)\mathbb{E}[\epsilon_k|\gamma_k]}{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)^2} \tag{32}$$

$$= \mathbb{E}\beta_k \tag{33}$$

Result (1) trivially follows the same logic.

□

G 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

$$\mathbb{E}\gamma_k^2 = \sigma_\gamma^2 + \mu_\gamma^2 \tag{34}$$

$$\mathbb{E}2\gamma_k u_k = 0 \text{ by } Cov(\gamma_k, u_k) = 0 \tag{35}$$

$$\mathbb{E}u_k^2 = \sigma_{uk}^2 \tag{36}$$

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

Now, considering the estimator,

$$\hat{\beta} = \frac{\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k}{\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{37}$$

$$= \frac{\frac{1}{K}\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})[\mathbb{E}\delta + \gamma_k\beta_k + (\varepsilon_k + v_k)]}{\frac{1}{K}\sum_{k=1}^K (\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{38}$$

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

$$\hat{\beta} = \frac{\frac{1}{K}\sum_{k=1}^K(\hat{\gamma}_k - \overline{\hat{\gamma}_k})[\mathbb{E}\delta + \gamma_k\beta_k + (\varepsilon_k + v_k)]}{\frac{1}{K}\sum_{k=1}^K(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{39}$$

$$\xrightarrow{p} \lambda\mathbb{E}\beta_k. \tag{40}$$

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

□

H 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 D:

$$\hat{\beta}_{ideal} = \frac{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)\tau_k}{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)^2} \tag{41}$$

$$\to \frac{\sigma_\gamma^2 \mathbb{E}\beta_k}{\sigma_\gamma^2} \tag{42}$$

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

$$\hat{\beta} = \frac{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k}{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{43}$$

$$= \frac{\frac{1}{K}\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})[\mathbb{E}\delta + \gamma_k\beta_k + (\varepsilon_k + v_k)]}{\frac{1}{K}\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{44}$$

$$\to \frac{\sigma_\gamma^2 \mathbb{E}\beta_k}{\sigma_\gamma^2 + \overline{\sigma_{uk}^2}} \tag{45}$$

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

$$\hat{\beta}_{BCES} = \frac{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k}{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2 - \sum_{k=1}^{K}\sigma_{uk}^2} \tag{46}$$

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:

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

I Proof of Proposition 3

Proof. We start from simplifying some notations. The overall average indirect effect is $\mathbb{E}[\eta_k] = \mathbb{E}[\beta_k]\sum_{k=1}^{K}\gamma_k\mathbb{P}[\gamma_k]$. We use λ_k to simplify the notation $\mathbb{P}[\gamma_k]$, the proportion of γ_k in the population, and then we use γ_0 to denote $\sum_{k=1}^{K}\gamma_k\lambda_k$. Under simplified notation, the overall average indirect effect is $\mathbb{E}[\beta_k]\gamma_0$. It is estimated by $\hat{\beta}\hat{\lambda}_0 := \hat{\beta}\sum_{k=1}^{K}\gamma_k\hat{\lambda}_k$, where we let $\hat{\lambda}_k$ to denote the consistent estimate of the proportion λ_k, and $\hat{\lambda}_0$ denotes $\sum_{k=1}^{K}\gamma_k\hat{\lambda}_k$.

The null hypothesis we consider is $H_0 : \mathbb{E}[\beta_k]\gamma_0 = 0$, which can be expressed as the union, $H_0 : \mathbb{E}[\beta_k] = 0 \cup \lambda_0 = 0$. Therefore, H_0 is rejected if we reject both parts of the null hypothesis. It is known that the (asymptotic) level α test is given by the (asymptotic) level α test of both parts (for example see Berger and Casella 2001, theorem 8.3.23).

To test whether $\mathbb{E}[\beta_k] = 0$, we apply the asymptotic normality of $\hat{\beta}_{SIMEX}$ by Carroll et al. (1996). Therefore, as traditional test, the p-value is $2 * \Phi_Z(-|\frac{\hat{\beta}_{SIMEX}}{\hat{\sigma}_{SIMEX}}|)$, which is p_β.

To test $\lambda_0 = 0$, given normality assumption, we know $\hat{\lambda}_0 \to N(\lambda_0, \sum_{k=1}^{K}\lambda^2\sigma_{vk}^2)$. Therefore, as traditional test, the p-value is $2 * \Phi_Z(-|\frac{\hat{\gamma}_0}{\sum_{k=1}^{K}\hat{\lambda}^2\sigma_{vk}^2}|)$, which is p_γ. It follows that rejection of the Null hypothesis if $p_\beta \leq \alpha$ and $p_\gamma \leq \alpha$ is an asymptotic level α test. □

J Proof of Proposition 4

Proof. Let $\Phi_Z(\cdot) := \Phi_Z(\frac{1+\sqrt{1-\alpha}}{2})$. From the proof of 3, we can construct the separate asymptotic $\sqrt{(1-\alpha)}$% confidence interval as

$$\mathbb{P}[\hat{\lambda}_0 - \Phi_Z(\cdot) * \sum_{k=1}^{K}\hat{\lambda}^2\sigma_{vk}^2 \leq \gamma_0 \leq \hat{\lambda}_0 + \Phi_Z(\cdot) * \sum_{k=1}^{K}\hat{\lambda}^2\sigma_{vk}^2] = \sqrt{(1-\alpha)}$$

$$\mathbb{P}[\hat{\beta}_{SIMEX} - \Phi_Z(\cdot) * \hat{\sigma}_{SIMEX} \leq \mathbb{E}[\beta_k] \leq \hat{\beta}_{SIMEX} + \Phi_Z(\cdot) * \hat{\sigma}_{SIMEX}] = \sqrt{(1-\alpha)}$$

that is,

$$\mathbb{P}[a_1 \leq \gamma_0 \leq a_2] = \sqrt{(1-\alpha)}$$
$$\mathbb{P}[a_3 \leq \mathbb{E}[\beta_k] \leq a_4] = \sqrt{(1-\alpha)}$$

Because $a_1 \leq \gamma_0 \leq a_2$ and $a_3 \leq \mathbb{E}[\beta_k] \leq a_4$ implies $\min[a_j] \leq \mathbb{E}[\beta_k]\gamma_0 \leq \max[a_j], j \in \{1,2,3,4\}$, we get

$$\mathbb{P}[\underline{a} \leq \mathbb{E}[\beta_k]\gamma_0 \leq \overline{a}] \geq 1-\alpha$$

□

K More on Simulation and Power analysis

K.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.

Algorithm 1 Monte Carlo Simulation

Input: : $k > 0$, β, $\delta = 4$

$\gamma \sim Norm(2,1)$, $\epsilon \sim Norm(0,1)$,
$\tau_k = \delta_k + \beta\gamma_k + \epsilon$
$\sigma_{uk} \sim Gamma(1,1)$, $\sigma_{vk} \sim Gamma(1,1)$
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: $\{\hat{\beta}_b\}$

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 A.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 estimator, 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

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 A.1: Monte Carlo Simulation and Rejection rate.

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.

Therefore, we strongly recommend the SIMEX estimator in real application. In particular, when BCES and SIMEX estimators have different signs, we should accpet the latter one because we find that BCES is pretty unstable.

K.2 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).

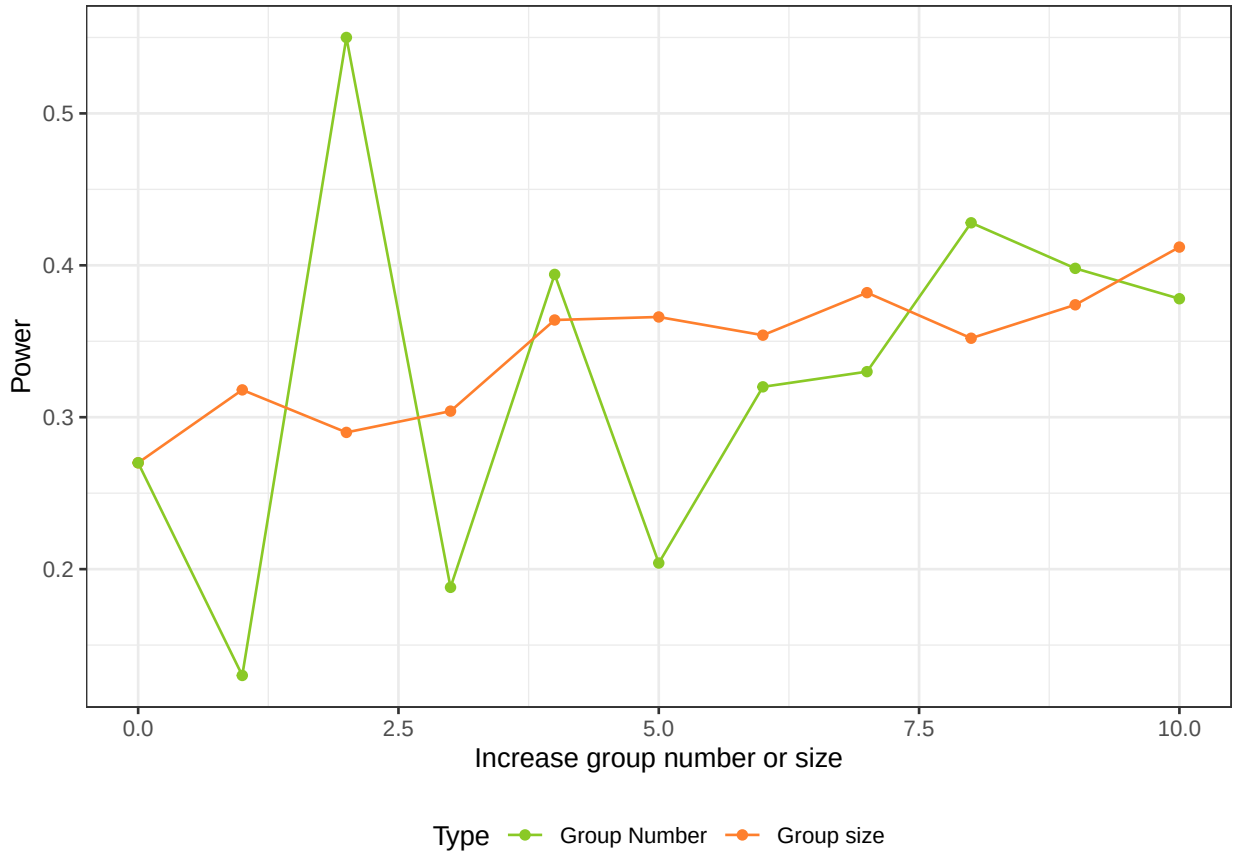

Figure A.1: Power analysis: Group number and size

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

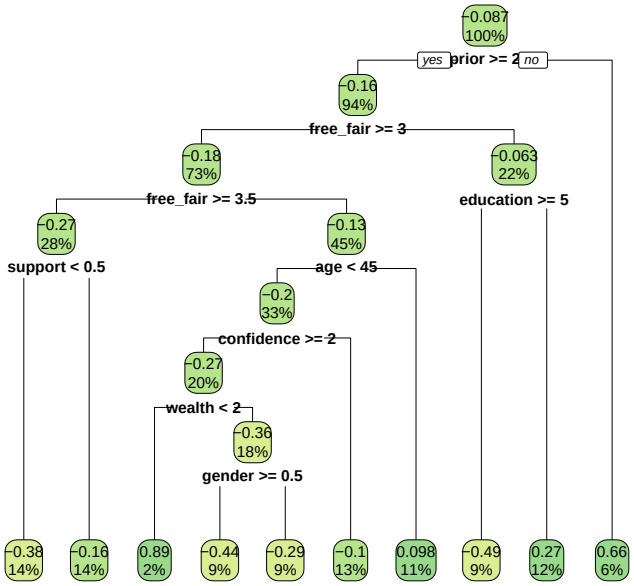

Figure A.2: Heterogeneous Subgroup Design

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.

K.3 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.

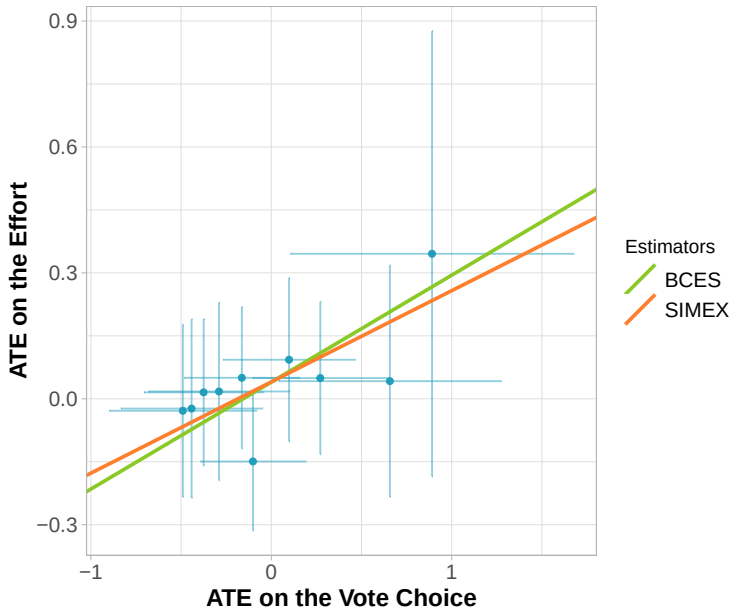

Figure A.3: Heterogeneous Subgroup Design