Supplementary Information

A Multiple Mechanisms

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

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

The direct effect is defined as

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

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

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

The overall causal effect can be decomposed as:

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

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

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

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

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

$$\tau^i = (\delta^i + \sum_{j=2}^{J} \eta_j^i) + \eta_1^i \tag{29}$$

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

B Proof of Proposition 1

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

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

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

$$\tau_k = \mathbb{E}\delta_k + \beta_k\gamma_k + \varepsilon_k \tag{30}$$

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

$$\hat{\beta} = \frac{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)\tau_k}{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} \tag{31}$$

$$= \frac{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)(\overline{\delta}_k + \beta_k\gamma_k + \varepsilon_k)}{\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} \tag{32}$$

$$= \frac{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)\beta_k\gamma_k}{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} + \frac{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)\varepsilon_k}{\frac{1}{K}\sum_{k=1}^K(\gamma_k - \overline{\gamma}_k)^2} \tag{33}$$

$$\xrightarrow{p} \frac{\mathbb{E}[(\gamma_k - \overline{\gamma}_k)\gamma_k\beta_k]}{Var(\gamma_k)} + \frac{\mathbb{E}[(\gamma_k - \overline{\gamma}_k)\varepsilon_k]}{Var(\gamma_k)} \tag{34}$$

$$= \frac{\mathbb{E}[(\gamma_k - \overline{\gamma}_k)\gamma_k]\mathbb{E}\beta_k}{Var(\gamma_k)} + \frac{\mathbb{E}\gamma_k\varepsilon_k}{Var(\gamma_k)} \tag{35}$$

$$= \mathbb{E}\beta_k \tag{36}$$

where line (33) comes from $\bar{\delta}_k \sum_{k=1}^{K}(\gamma_k - \bar{\gamma}_k) = 0$, line (34) is implied by Slutsky's Lemma, (35) is implied by $\mathbb{E}\varepsilon_k$=0 and assumption $\beta_k \perp\!\!\!\perp \gamma_k$, the last line is implied by $\mathbb{E}[\gamma_k \epsilon_k] = 0$,

Result (1) trivially follows the same logic.

□

C Proposition 1 under Multiple Mechanisms

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

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

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

$$\tau = \Delta + \eta_1 \tag{37}$$

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

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

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

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

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

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

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

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

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

D Proof of Unbiasedness

Proposition 4. *Let* (τ, δ, γ) *be random variables and as defined in* (11) *and* (12). *Given the random sample* $(\tau_k, \gamma_k)_{k \in K}$. *Suppose following two assumptions hold:*

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

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

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

(1) If β *is a constant, then* $\mathbb{E}\hat{\beta} = \beta$;

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

Proof. For unbiasedness, note that by construction, $\mathbb{E}\varepsilon_k = \mathbb{E}\delta_k - \mathbb{E}\delta_k = 0$ and thus with mean independence assumption (2) we have $\mathbb{E}[\epsilon_k|\gamma_k] = \mathbb{E}\varepsilon_k = 0$. From line (33), we take the expectation given observed $\gamma_1, \gamma_2, ..., \gamma_K$,

$$\mathbb{E}[\hat{\beta}|\gamma_1, \gamma_2, ..., \gamma_K] = \mathbb{E}[\beta_k]\frac{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)\gamma_k}{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)^2} + \frac{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)\mathbb{E}[\epsilon_k|\gamma_k]}{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma}_k)^2} \tag{38}$$

$$= \mathbb{E}\beta_k \tag{39}$$

Result (1) trivially follows the same logic.

□

E Proof of Proposition 2

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

For each part, we have

$$\mathbb{E}\gamma_k^2 = \sigma_\gamma^2 + \mu_\gamma^2 \tag{40}$$

$$\mathbb{E}2\gamma_k u_k = 0 \text{ by } Cov(\gamma_k, u_k) = 0 \tag{41}$$

$$\mathbb{E}u_k^2 = \sigma_{uk}^2 \tag{42}$$

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

Now, considering the estimator,

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

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

To 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

$$\hat{\beta} = \frac{\frac{1}{K}\sum_{k=1}^K(\hat{\gamma}_k - \overline{\hat{\gamma}_k})[\mathbb{E}\delta + \gamma_k\beta_k + (\varepsilon_k + v_k)]}{\frac{1}{K}\sum_{k=1}^K(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{45}$$

$$\xrightarrow{p} \lambda\mathbb{E}\beta_k. \tag{46}$$

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

□

F BCES estimator

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

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

$$\to \frac{\sigma_\gamma^2 \mathbb{E}\beta_k}{\sigma_\gamma^2} \tag{48}$$

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

$$\hat{\beta} = \frac{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k}{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{49}$$

$$= \frac{\frac{1}{K}\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})[\mathbb{E}\delta + \gamma_k\beta_k + (\varepsilon_k + v_k)]}{\frac{1}{K}\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2} \tag{50}$$

$$\to \frac{\sigma_\gamma^2 \mathbb{E}\beta_k}{\sigma_\gamma^2 + \overline{\sigma_{uk}^2}} \tag{51}$$

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

$$\hat{\beta}_{BCES} = \frac{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k}{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2 - \sum_{k=1}^{K}\sigma_{uk}^2} \tag{52}$$

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:

$$\hat{\beta}_{BCES} = \frac{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})\hat{\tau}_k - \sum_{k=1}^{K}\sigma_{uvk}^2}{\sum_{k=1}^{K}(\hat{\gamma}_k - \overline{\hat{\gamma}_k})^2 - \sum_{k=1}^{K}\sigma_{uk}^2} \tag{53}$$

G Details in the Bootstrap

Restricted Wild Bootstrap

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

Pairs Bootstrap

The same as above except Step 3 (a):

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

H Power analysis

We investigate how the power changes with the number of groups and the number of individuals in 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 $k * n$ people according to the budget (k is an integer). They have to decide to add one more group (i.e. let additional $k * n$ individuals form k new group) or increase the group size (i.e. add $\frac{kn}{10}$ individuals to each existing groups).

The data generating process is the same as the one in the main text. When increasing group size, we need increase the precision of the observed values. To do so, we apply the following approximation:

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

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

In the figure A.1, the vertical line the power. The number on the horizontal line is k, which denotes k more groups (green line) or $kn/10$ more individuals in each group (orange line).

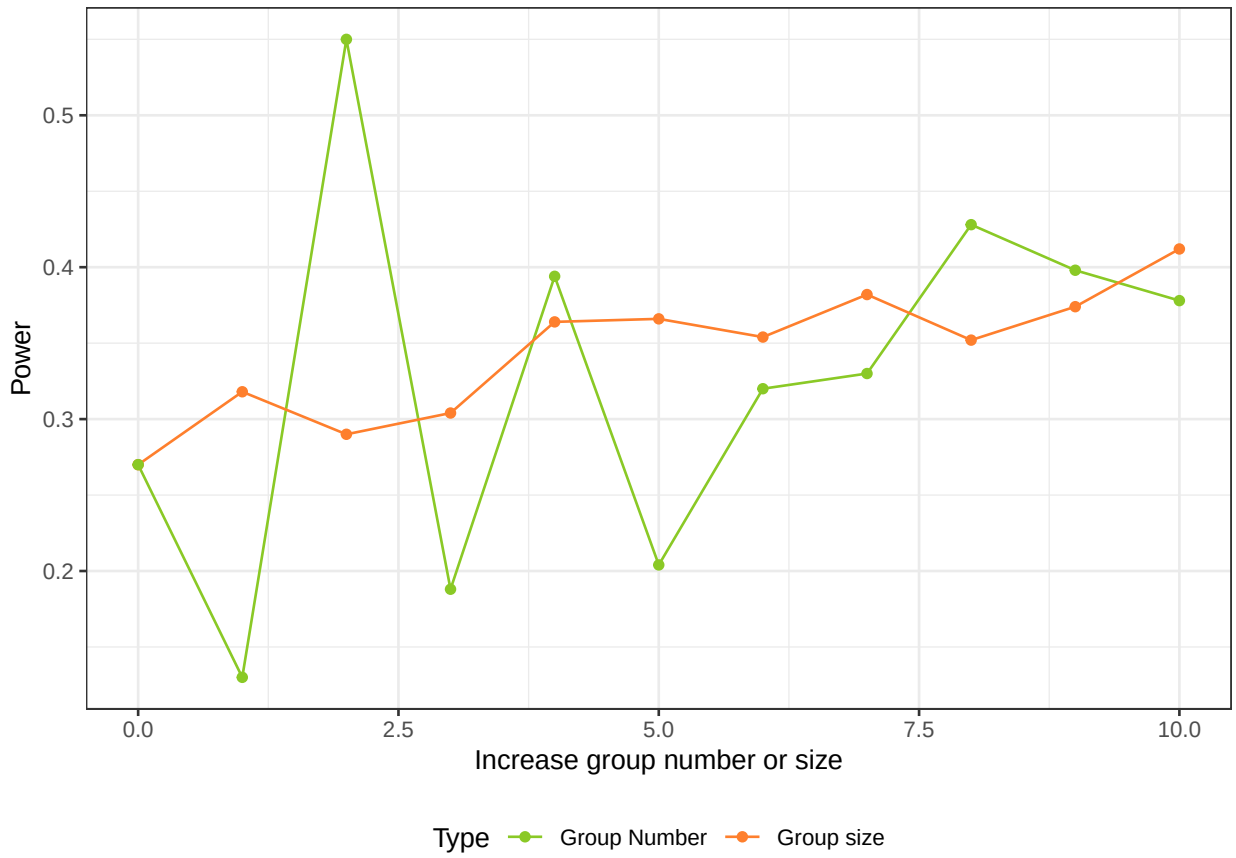

Figure A.1: Power analysis: Group number and size

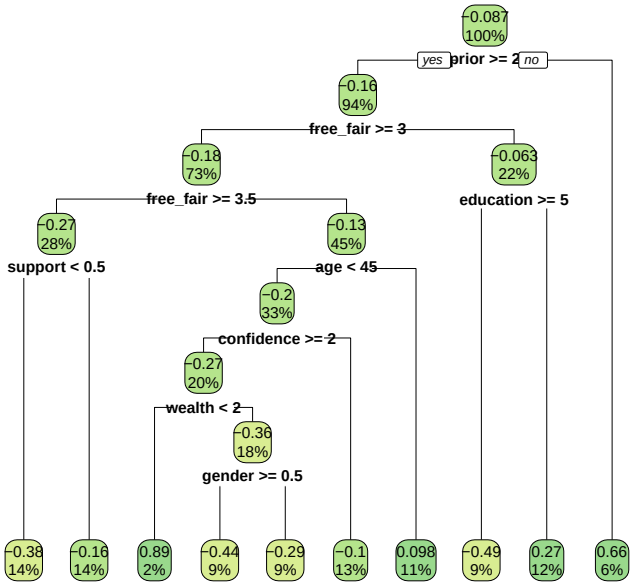

Figure A.2: Heterogeneous Subgroup Design

I Application II

In this section, we use full data to find subgroups with causal trees. As shown in figure A.2, more subgroups are detected. Actually, the number of groups is affected by many factors, such as the minimal number of observations allowed in each split. The corresponding estimates are shown in figure A.3. As we can see, the estimate of β is similar to the one in the main text.

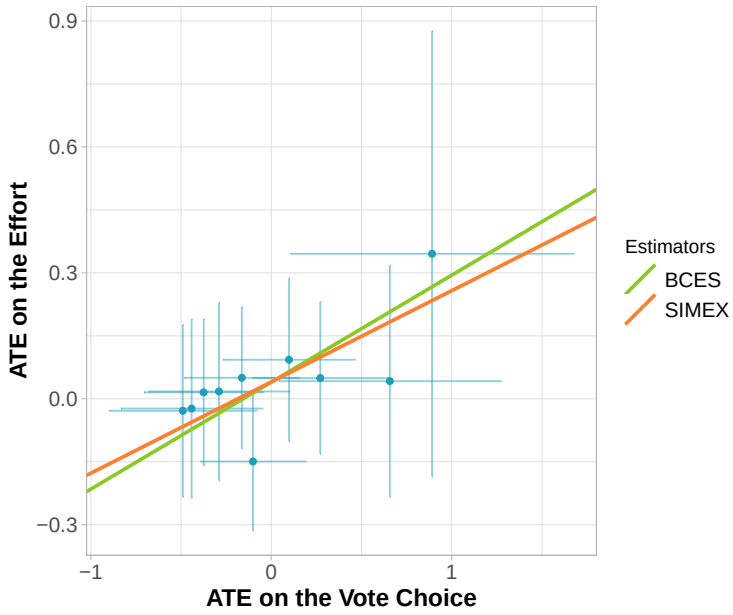

Figure A.3: Heterogeneous Subgroup Design