Improving LATE Estimation in Experiments with Imperfect Compliance

2023 EEA-ESEM Congress

Yagan Hazard, Simon Löwe

August 29, 2023

Motivation

Imperfect compliance in causal effect estimation

2 important consequences of imperfect compliance

1. Identification of a Local Average Treatment Effect (LATE)
2. Low avr. compliance $\Rightarrow$ possibly uninformative inference
   (high variance in estimation)

Motivation

Imperfect compliance in causal effect estimation

2 important consequences of imperfect compliance

1. Identification of a Local Average Treatment Effect (LATE)
2. Low avr. compliance $\Rightarrow$ possibly uninformative inference
   (high variance in estimation)

$$\text{Var}\left[\widehat{2SLS}\right] = \frac{1}{N} \cdot \frac{1}{\pi^2} \cdot \frac{\sigma_\varepsilon^2}{p \cdot (1-p)}$$

N: sample size, p: sh. encouraged indiv., π: sh. compliers, σ_ε: sd. errors

Motivation

Imperfect compliance in causal effect estimation

2 important consequences of imperfect compliance

1. Identification of a Local Average Treatment Effect (LATE)
2. Low avr. compliance $\Rightarrow$ possibly uninformative inference
   (high variance in estimation)

$$\text{Var}\left[\widehat{2SLS}\right] = \frac{1}{N} \cdot \frac{1}{\pi^2} \cdot \frac{\sigma_\varepsilon^2}{p \cdot (1-p)}$$

N: sample size, p: sh. encouraged indiv., π: sh. compliers, σ_ε: sd. errors

Let's assume: $\text{P}[G=0] = 0.5, \pi^{G=0} = 0$ and $\text{P}[G=1] = 0.5, \pi^{G=1} = 0.10$

$$(N_a = 10,000\ ;\ \pi_a = 0.05) \quad \rightarrow \quad (N_b = 5,000\ ;\ \pi_b = 0.10)$$

$$\frac{1}{N_a} \cdot \frac{1}{\pi_a^2} \cdot \frac{\sigma_\varepsilon^2}{p \cdot (1-p)} \quad \rightarrow \quad \frac{1}{N_a/2} \cdot \frac{1}{(\pi_a \times 2)^2} \cdot \frac{\sigma_\varepsilon^2}{p \cdot (1-p)}$$

$$= \frac{1}{2} \cdot \left(\frac{1}{N_a} \cdot \frac{1}{\pi_a^2} \cdot \frac{\sigma_\varepsilon^2}{p \cdot (1-p)} \right)$$

This paper

From Insight to Practice

Illustrative example suggests potential gains in precision in case of heterogeneous first-stage along observables.

Yet in practice, no *ex-ante* knowledge of which pop. does not comply...

How can we still exploit such first-stage heterogeneity?

This paper

From Insight to Practice

Illustrative example suggests potential gains in precision in case of heterogeneous first-stage along observables.

Yet in practice, no *ex-ante* knowledge of which pop. does not comply...

How can we still exploit such first-stage heterogeneity?

Intuitive ("naïve") Test-and-Select procedure:

(i) t-test first-stage coef. (compliance rate) by group

(ii) restrict 2SLS estimation to groups with significant first-stages

$\rightarrow$ would yield a biased estimator.

This paper

Preview of the results

1. Study of the bias of the “naïve” selection rule

This paper

Preview of the results

1. Study of the bias of the “naïve” selection rule
2. Proposition of a “sophisticated” selection rule to correct bias

   $\rightarrow$ Test-and-Select (TS henceforth) estimator

This paper

Preview of the results

1. Study of the bias of the "naïve" selection rule
2. Proposition of a "sophisticated" selection rule to correct bias

   $\rightarrow$ Test-and-Select (TS henceforth) estimator
3. Result 1: under "standard" asymptotic sequences, TS estimator is shown to be asymptotically normal and unbiased for the LATE

This paper

Preview of the results

1. Study of the bias of the "naïve" selection rule
2. Proposition of a "sophisticated" selection rule to correct bias

   $\rightarrow$ Test-and-Select (TS henceforth) estimator
3. Result 1: under "standard" asymptotic sequences, TS estimator is shown to be asymptotically normal and unbiased for the LATE

Henceforth, $\hat{\theta}$ (estimator of θ) is said to be asymp. normal and unbiased if

$$\sqrt{n} \cdot \left(\hat{\theta} - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Sigma)$$

It has a "first-order" or "asymptotic" bias B if

$$\sqrt{n} \cdot \left(\hat{\theta} - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(B, \Sigma)$$

$\rightarrow$ can lead to invalid inference (CIs centered on $\theta + B$ instead of θ).

This paper

Preview of the results

1. Study of the bias of the "naïve" selection rule
2. Proposition of a "sophisticated" selection rule to correct bias

   $\rightarrow$ Test-and-Select (TS henceforth) estimator
3. Result 1: under "standard" asymptotic sequences, TS estimator is shown to be asymptotically normal and unbiased for the LATE
4. Result 2: under "weak-IV-like" asymptotics (better approx. of finite sample behavior) , TS estimator asymptotically normal and unbiased for the LATE under restrictions on TE heterogeneity

   $\rightarrow$ compared to alternative estimators exploiting 1^{st}-stage het.:

   $\sqrt{n}$-consistent under less restrictions on treatment effect heterogeneity

This paper

Preview of the results

1. Study of the bias of the "naïve" selection rule
2. Proposition of a "sophisticated" selection rule to correct bias

   → Test-and-Select (TS henceforth) estimator
3. Result 1: under "standard" asymptotic sequences, TS estimator is shown to be asymptotically normal and unbiased for the LATE
4. Result 2: under "weak-IV-like" asymptotics (better approx. of finite sample behavior) , TS estimator asymptotically normal and unbiased for the LATE under restrictions on TE heterogeneity
5. Result 3: finite sample properties studied in Monte-Carlo simulations

This paper

Preview of the results

1. Study of the bias of the "naïve" selection rule
2. Proposition of a "sophisticated" selection rule to correct bias
   → Test-and-Select (TS henceforth) estimator
3. Result 1: under "standard" asymptotic sequences, TS estimator is shown to be asymptotically normal and unbiased for the LATE
4. Result 2: under "weak-IV-like" asymptotics (better approx. of finite sample behavior) , TS estimator asymptotically normal and unbiased for the LATE under restrictions on TE heterogeneity
5. Result 3: finite sample properties studied in Monte-Carlo simulations
6. Applications to a natural experiment and an encouragement design

Roadmap

Framework

Proposed estimator

Asymptotic results under “standard” sequences

Asymptotic results under “local-to-zero” sequences

Finite-sample properties: Monte-Carlo simulations

Practical guidance and Conclusion

Framework

Following Angrist, Imbens and Rubin (1996)

Data-generating process:

- Super-population:
  $(Y(1), Y(0), D(1), D(0), G, Z)$
- I.i.d. sample:
  $\{(Y_i(1), Y_i(0), D_i(1), D_i(0), G_i, Z_i)\}_{i=1}^{n}$
- Potential outcomes framework:
  $$Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$$
  $$D = Z \cdot D(1) + (1 - Z) \cdot D(0)$$

Assumption 1 (LATE identification)

1. Independence:
   $(Y(1), Y(0), D(1), D(0), G) \perp Z$
2. Exclusion restriction:
   $Y(D, Z) = Y(D)$
3. First Stage:
   $E[D(1) - D(0)] > c > 0, \;\; c \in \mathrm{R}^{+} \backslash \{0\}$
4. Monotonicity: $D(1) \geq D(0)$

Identified estimand: $\mathrm{LATE} \equiv \mathrm{E}[Y(1) - Y(0) | D(1) > D(0)]$

Framework

First-stage heterogeneity

Key addition to usual framework:

- Pre-determined discrete covariate $G \perp Z$

  [Rationale: quite common to collect additional *pre-experiment* covariates even in completely randomized experiments.]

- 1^{st} stage can be heterogeneous across groups

$$\pi^g \equiv E[D(1) - D(0)|G = g]$$

Framework

First-stage heterogeneity

Key addition to usual framework:

- Pre-determined discrete covariate $G \perp Z$

  [Rationale: quite common to collect additional *pre-experiment* covariates even in completely randomized experiments.]

- 1^{st} stage can be heterogeneous across groups

$$\pi^g \equiv E[D(1) - D(0)|G = g]$$

In practice, we do not know ex-ante for which g $\pi^g = 0$ vs. $\pi^g > 0$

$\Rightarrow$ we need to learn it from the data (testing).

What are the consequences of this pre-testing step?

How to best implement it in order to avoid bias?

Roadmap

Framework

Proposed estimator

Asymptotic results under “standard” sequences

Asymptotic results under “local-to-zero” sequences

Finite-sample properties: Monte-Carlo simulations

Practical guidance and Conclusion

“Naïve” Test-and-Select estimator

Procedure

Naïve Test-and-Select (Naïve TS) estimation procedure:

1. For each group $G = g$: one-sided t-test on $\hat{\pi}^g$ at level α
2. Select only groups for which we reject the null of $\pi^g = 0$

   Example of decision rule: t-stat$(\hat{\pi}^g) > 1.64$ $(\alpha = 0.95)$
3. Compute usual Wald (=2SLS) estimator on selected sample

Intuition: we drop only groups with no compliers *or* very few compliers, so we might not be too far away from the LATE.

“Naïve” Test-and-Select estimator

Procedure

Naïve Test-and-Select (Naïve TS) estimation procedure:

1. For each group $G = g$: one-sided t-test on $\hat{\pi}^g$ at level α
2. Select only groups for which we reject the null of $\pi^g = 0$

   Example of decision rule: t-stat$(\hat{\pi}^g) > 1.64$ $(\alpha = 0.95)$
3. Compute usual Wald (=2SLS) estimator on selected sample

Intuition: we drop only groups with no compliers *or* very few compliers, so we might not be too far away from the LATE.

Table: Bias introduced by naïve pre-test (Monte-Carlo simulations)

	2SLS	Naïve TS
Bias	0.003	-0.221
Coverage	0.953	0.861

Test-and-Select estimator

Procedure

Test-and-Select (TS) estimation procedure:

1. Divide sample in 2 equally sized random sub-samples $\mathcal{I}_1$ and $\mathcal{I}_2$

Test-and-Select estimator

Procedure

Test-and-Select (TS) estimation procedure:

1. Divide sample in 2 equally sized random sub-samples $\mathcal{I}_1$ and $\mathcal{I}_2$
   (stratifying the random split by G)
2. In $\mathcal{I}_1$: for each group $G = g$, one-sided t-test on $\hat{\pi}^g$ at a given level α

Test-and-Select estimator

Procedure

Test-and-Select (TS) estimation procedure:

1. Divide sample in 2 equally sized random sub-samples $\mathcal{I}_1$ and $\mathcal{I}_2$
   (stratifying the random split by G)
2. In $\mathcal{I}_1$: for each group $G = g$, one-sided t-test on $\hat{\pi}^g$ at a given level α
3. In $\mathcal{I}_2$: select groups for which the null of $\pi^g = 0$ was rejected in $\mathcal{I}_1$

Test-and-Select estimator

Procedure

Test-and-Select (TS) estimation procedure:

1. Divide sample in 2 equally sized random sub-samples $\mathcal{I}_1$ and $\mathcal{I}_2$ (stratifying the random split by G)
2. In $\mathcal{I}_1$: for each group $G = g$, one-sided t-test on $\hat{\pi}^g$ at a given level α
3. In $\mathcal{I}_2$: select groups for which the null of $\pi^g = 0$ was rejected in $\mathcal{I}_1$
4. Compute 2SLS estimator on the selected sub-sample of $\mathcal{I}_2$.

Test-and-Select estimator

Procedure

Test-and-Select (TS) estimation procedure:

1. Divide sample in 2 equally sized random sub-samples $\mathcal{I}_1$ and $\mathcal{I}_2$ (stratifying the random split by G)
2. In $\mathcal{I}_1$: for each group $G = g$, one-sided t-test on $\hat{\pi}^g$ at a given level α
3. In $\mathcal{I}_2$: select groups for which the null of $\pi^g = 0$ was rejected in $\mathcal{I}_1$
4. Compute 2SLS estimator on the selected sub-sample of $\mathcal{I}_2$.
5. Repeat steps 2 to 4 reversing the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ (cross-fitting).
6. Take the average of the estimators obtained in 4. within $\mathcal{I}_1$ and $\mathcal{I}_2$.

Test-and-Select estimator

Procedure

Test-and-Select (TS) estimation procedure:

1. Divide sample in 2 equally sized random sub-samples $\mathcal{I}_1$ and $\mathcal{I}_2$
   (stratifying the random split by G)
2. In $\mathcal{I}_1$: for each group $G = g$, one-sided t-test on $\hat{\pi}^g$ at a given level α
3. In $\mathcal{I}_2$: select groups for which the null of $\pi^g = 0$ was rejected in $\mathcal{I}_1$
4. Compute 2SLS estimator on the selected sub-sample of $\mathcal{I}_2$.
5. Repeat steps 2 to 4 reversing the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ (cross-fitting).
6. Take the average of the estimators obtained in 4. within $\mathcal{I}_1$ and $\mathcal{I}_2$.

Intuition: data-splitting $\Rightarrow$ testing step $\perp$ estimation step

$\Rightarrow$ no 1^{st}-order bias

cross-fitting $\Rightarrow$ reduces the efficiency loss due to splitting

(see Lemma 3 in working paper)

Test-and-Select estimator

Procedure

Test-and-Select (TS) estimation procedure:

1. Divide sample in 2 equally sized random sub-samples $\mathcal{I}_1$ and $\mathcal{I}_2$ (stratifying the random split by G)
2. In $\mathcal{I}_1$: for each group $G = g$, one-sided t-test on $\hat{\pi}^g$ at a given level α
3. In $\mathcal{I}_2$: select groups for which the null of $\pi^g = 0$ was rejected in $\mathcal{I}_1$
4. Compute 2SLS estimator on the selected sub-sample of $\mathcal{I}_2$.
5. Repeat steps 2 to 4 reversing the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ (cross-fitting).
6. Take the average of the estimators obtained in 4. within $\mathcal{I}_1$ and $\mathcal{I}_2$.

Table: 1^{st}-order bias correction (Monte-Carlo simulations)

	2SLS	Naïve TS	TS
Bias	0.003	-0.221	0.097
Coverage	0.953	0.861	0.976

Roadmap

Framework

Proposed estimator

Asymptotic results under “standard” sequences

Asymptotic results under “local-to-zero” sequences

Finite-sample properties: Monte-Carlo simulations

Practical guidance and Conclusion

Framework

"Standard" asymptotics

Assumption 2 (No weak 1^{st}-stages)

For any group $g \in |\mathcal{G}|$, we have either $\pi^g = 0$ or $\pi^g \geq c,\ c \in \mathbb{R}^+ \backslash \{0\}$.
In other words: there are only groups with "zero" or "strong" 1^{st}-stages.

Let us stick to a 2-group case ($G \in \{0, 1\}$) for clarity of exposition. Our assumptions (1 and 2) imply:

- LATE is identified: overall first-stage well separated from 0

$$\pi \equiv \pi^0 \cdot \mathrm{P}[G = 0] + \pi^1 \cdot \mathrm{P}[G = 1] > c' > 0$$

- One of the conditional LATEs might be unidentified, e.g.:

$$\left(\pi^0 = 0,\ \pi^1 > c\right) \quad \text{or} \quad \left(\pi^0 > c,\ \pi^1 = 0\right) \quad \text{or} \quad \left(\pi^0 > c,\ \pi^1 > c\right)$$

Let's consider (wlog) the case: $\left(\pi^0 = 0,\ \pi^1 > c\right)$

Limiting distribution of the TS estimator

under "standard" asymptotics

Recall we have 2 sub-samples: $\mathcal{I}_1$ (testing sample) and $\mathcal{I}_2$ (estimation sample)

Selection rule determined in $\mathcal{I}_1$. We get (asymptotically):

$$P[\{G = 1 \text{ is selected}\}] \xrightarrow[n\to\infty]{} 1$$

$$P[\{G = 0 \text{ is selected}\}] \xrightarrow[n\to\infty]{} \alpha$$

Hence in $\mathcal{I}_2$, we compute:

$$\text{w/ proba. } \alpha: \quad \widehat{\text{TS}}_n = \widehat{\text{2SLS}}_n = \frac{\text{E}_n[Y|Z=1] - \text{E}_n[Y|Z=0]}{\text{E}_n[D|Z=1] - \text{E}_n[D|Z=0]}$$

$$\sqrt{n} \cdot (\widehat{\text{2SLS}}_n - LATE) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS})$$

Variance gains

Limiting distribution of the TS estimator

under "standard" asymptotics

Recall we have 2 sub-samples: $\mathcal{I}_1$ (testing sample) and $\mathcal{I}_2$ (estimation sample)

Selection rule determined in $\mathcal{I}_1$. We get (asymptotically):

$$P[\{G = 1 \text{ is selected}\}] \xrightarrow[n\to\infty]{} 1$$

$$P[\{G = 0 \text{ is selected}\}] \xrightarrow[n\to\infty]{} \alpha$$

Hence in $\mathcal{I}_2$, we compute:

$$\text{w/ proba. } \alpha: \quad \widehat{\text{TS}}_n = \widehat{\text{2SLS}}_n = \frac{\text{E}_n[Y|Z=1] - \text{E}_n[Y|Z=0]}{\text{E}_n[D|Z=1] - \text{E}_n[D|Z=0]}$$

$$\sqrt{n} \cdot (\widehat{\text{2SLS}}_n - LATE) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS})$$

$$\text{w/ proba. } (1-\alpha): \quad \widehat{\text{TS}}_n = \widehat{\text{2SLS}^1}_n = \frac{\text{E}_n[Y|Z=1, G=1] - \text{E}_n[Y|Z=0, G=1]}{\text{E}_n[D|Z=1, G=1] - \text{E}_n[D|Z=0, G=1]}$$

$$\sqrt{n} \cdot (\widehat{\text{2SLS}}^1_n - \underbrace{LATE^1}_{=LATE}) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \underbrace{V^{2SLS^1}}_{\leq V^{2SLS}})$$

Variance gains

Summary of results under “standard” asymptotics

So far:

- $\widehat{TS}$ is asymptotically normal and unbiased for the LATE...
- ... w/ smaller variance when selection occurs $\Rightarrow$ improved inference

(shorter CIs)

Summary of results under “standard” asymptotics

So far:

- $\widehat{TS}$ is asymptotically normal and unbiased for the LATE...
- ... w/ smaller variance when selection occurs $\Rightarrow$ improved inference

(shorter CIs)

Too good to be true?

- In finite samples: we expect groups w/ small share of compliers to be dropped in the process $\rightarrow$ not captured by our asymptotic approx.
- How much bias would that create?

Summary of results under "standard" asymptotics

So far:

- $\widehat{TS}$ is asymptotically normal and unbiased for the LATE...
- ... w/ smaller variance when selection occurs $\Rightarrow$ improved inference (shorter CIs)

Too good to be true?

- In finite samples: we expect groups w/ small share of compliers to be dropped in the process $\rightarrow$ not captured by our asymptotic approx.
- How much bias would that create?

$\Rightarrow$ Change asymptotic sequence for credible finite sample approx.

Roadmap

Framework

Proposed estimator

Asymptotic results under “standard” sequences

Asymptotic results under “local-to-zero” sequences

Finite-sample properties: Monte-Carlo simulations

Practical guidance and Conclusion

Framework

"Local-to-zero" asymptotics

Assumption 3 (Some weak 1st-stages)

Some groups have "local-to-zero" (Staiger and Stock, 1997) share of compliers.

Formally: $$\exists g \in \mathcal{G} \text{ s.t. } \pi^g = \frac{H^g}{\sqrt{n}}, \text{ with } H^g \in \mathbb{R}^+ \backslash \{0\}$$

Why would this yield a better approximation of finite sample behavior?

$\rightarrow$ Because it allows for de-selection of groups with non-zero shares of compliers asymptotically.

Framework

"Local-to-zero" asymptotics

Assumption 3 (Some weak 1st-stages)

Some groups have "local-to-zero" (Staiger and Stock, 1997) share of compliers.

Formally: $$\exists g \in \mathcal{G} \text{ s.t. } \pi^g = \frac{H^g}{\sqrt{n}}, \text{ with } H^g \in \mathbb{R}^+ \backslash \{0\}$$

Again, let us stick to a 2-group case ($G \in \{0, 1\}$) – a "weak" and a "strong" one. Assumptions 1 and 3 imply:

- LATE is identified: overall first-stage well separated from 0

$$\pi \equiv \pi^0 \cdot \mathrm{P}[G = 0] + \pi^1 \cdot \mathrm{P}[G = 1] > c' > 0$$

- One of the conditional LATEs might be weakly identified, e.g.:

$$\left(\pi^0 = \frac{H}{\sqrt{n_0}},\ \pi^1 > c\right) \text{ or } \left(\pi^0 > c,\ \pi^1 = \frac{H}{\sqrt{n_1}}\right) \text{ or } \left(\pi^0 > c,\ \pi^1 > c\right)$$

Framework

"Local-to-zero" asymptotics

Assumption 3 (Some weak 1st-stages)

Some groups have "local-to-zero" (Staiger and Stock, 1997) share of compliers.

Formally: $$\exists g \in \mathcal{G} \text{ s.t. } \pi^g = \frac{H^g}{\sqrt{n}}, \text{ with } H^g \in \mathbb{R}^+ \backslash \{0\}$$

Again, let us stick to a 2-group case ($G \in \{0, 1\}$) – a "weak" and a "strong" one. Assumptions 1 and 3 imply:

- LATE is identified: overall first-stage well separated from 0

$$\pi \equiv \pi^0 \cdot \mathrm{P}[G = 0] + \pi^1 \cdot \mathrm{P}[G = 1] > c' > 0$$

- One of the conditional LATEs might be weakly identified, e.g.:

$$\left(\pi^0 = \frac{H}{\sqrt{n_0}},\ \pi^1 > c\right) \text{ or } \left(\pi^0 > c,\ \pi^1 = \frac{H}{\sqrt{n_1}}\right) \text{ or } \left(\pi^0 > c,\ \pi^1 > c\right)$$

Limiting distribution of the TS estimator

under "local-to-zero" asymptotics

Recall we have 2 sub-samples: $\mathcal{I}_1$ (testing sample) and $\mathcal{I}_2$ (estimation sample)

Selection rule determined in $\mathcal{I}_1$. We get (asymptotically):

$$P[\{G = 1 \text{ is selected}\}] \xrightarrow[n\to\infty]{} 1$$

$$P[\{G = 0 \text{ is selected}\}] \xrightarrow[n\to\infty]{} q \in (\alpha, 1)$$

Hence in $\mathcal{I}_2$, we compute:

$$\text{w/ proba. } q: \quad \widehat{\text{TS}}_n = \widehat{\text{2SLS}}_n = \frac{\text{E}_n[Y|Z=1] - \text{E}_n[Y|Z=0]}{\text{E}_n[D|Z=1] - \text{E}_n[D|Z=0]}$$

$$\sqrt{n} \cdot (\widehat{\text{2SLS}}_n - LATE) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS})$$

Limiting distribution of the TS estimator

under "local-to-zero" asymptotics

Recall we have 2 sub-samples: $\mathcal{I}_1$ (testing sample) and $\mathcal{I}_2$ (estimation sample)

Selection rule determined in $\mathcal{I}_1$. We get (asymptotically):

$$P[\{G=1 \text{ is selected}\}] \xrightarrow[n\to\infty]{} 1$$

$$P[\{G=0 \text{ is selected}\}] \xrightarrow[n\to\infty]{} q \in (\alpha, 1)$$

Hence in $\mathcal{I}_2$, we compute:

$$\text{w/ proba. } q: \quad \widehat{\mathrm{TS}}_n = \widehat{2\mathrm{SLS}}_n = \frac{\mathrm{E}_n[Y|Z=1] - \mathrm{E}_n[Y|Z=0]}{\mathrm{E}_n[D|Z=1] - \mathrm{E}_n[D|Z=0]}$$

$$\sqrt{n}\cdot(\widehat{2\mathrm{SLS}}_n - LATE) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS})$$

$$\text{w/ proba. } (1-q): \quad \widehat{\mathrm{TS}}_n = \widehat{2\mathrm{SLS}^1}_n = \frac{\mathrm{E}_n[Y|Z=1,G=1] - \mathrm{E}_n[Y|Z=0,G=1]}{\mathrm{E}_n[D|Z=1,G=1] - \mathrm{E}_n[D|Z=0,G=1]}$$

$$\sqrt{n}\cdot(\widehat{2\mathrm{SLS}}^1_n - \underbrace{LATE^1}_{\neq LATE}) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \underbrace{V^{2SLS^1}}_{\leq V^{2SLS}})$$

Limiting distribution of the TS estimator

under "local-to-zero" asymptotics

Recall we have 2 sub-samples: $\mathcal{I}_1$ (testing sample) and $\mathcal{I}_2$ (estimation sample)

Selection rule determined in $\mathcal{I}_1$. We get (asymptotically):

$$P[\{G = 1 \text{ is selected}\}] \xrightarrow[n\to\infty]{} 1$$

$$P[\{G = 0 \text{ is selected}\}] \xrightarrow[n\to\infty]{} q \in (\alpha, 1)$$

Hence in $\mathcal{I}_2$, we compute:

$$\text{w/ proba. } q: \quad \widehat{\text{TS}}_n = \widehat{\text{2SLS}}_n = \frac{\text{E}_n[Y|Z=1] - \text{E}_n[Y|Z=0]}{\text{E}_n[D|Z=1] - \text{E}_n[D|Z=0]}$$

$$\sqrt{n} \cdot (\widehat{\text{2SLS}}_n - LATE) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS})$$

$$\text{w/ proba. } (1-q): \quad \widehat{\text{TS}}_n = \widehat{\text{2SLS}^1}_n = \frac{\text{E}_n[Y|Z=1, G=1] - \text{E}_n[Y|Z=0, G=1]}{\text{E}_n[D|Z=1, G=1] - \text{E}_n[D|Z=0, G=1]}$$

$$\sqrt{n} \cdot (\widehat{\text{2SLS}}_n^1 - LATE^1) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS^1})$$

Asymptotic bias

under "local-to-zero" asymptotics

$$\sqrt{n} \cdot (\widehat{2\text{SLS}}_n^1 - \underbrace{LATE^1}_{\neq LATE}) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS^1})$$

Asymptotic bias

under "local-to-zero" asymptotics

$$\sqrt{n} \cdot (\widehat{\text{2SLS}}_n^1 - \underbrace{LATE^1}_{\neq LATE}) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS^1})$$

$$\underbrace{\sqrt{n} \cdot (\widehat{\text{2SLS}}_n^1 - LATE)}_{(a_n)} \xrightarrow[n\to\infty]{d} ?$$

Asymptotic bias

under "local-to-zero" asymptotics

$$\sqrt{n} \cdot (\widehat{\text{2SLS}}_n^1 - LATE^1) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS^1})$$

$$\underbrace{\sqrt{n} \cdot (\widehat{\text{2SLS}}_n^1 - LATE)}_{(a_n)} \xrightarrow[n\to\infty]{d} ?$$

$$(a_n) = \sqrt{n} \cdot (\widehat{\text{2SLS}}_n^1 - LATE^1) + \sqrt{n} \cdot (LATE^1 - LATE)$$

$$\xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS^1}) + B$$

where B is defined as: $\sqrt{n} \cdot \left(\text{LATE}^1 - \text{LATE}\right) \xrightarrow[n\to\infty]{} B$

Asymptotic bias

under "local-to-zero" asymptotics

$$\sqrt{n} \cdot (\widehat{\text{2SLS}}_n^1 - LATE^1) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS^1})$$

$$\underbrace{\sqrt{n} \cdot (\widehat{\text{2SLS}}_n^1 - LATE)}_{(a_n)} \xrightarrow[n\to\infty]{d} ?$$

$$(a_n) = \sqrt{n} \cdot (\widehat{\text{2SLS}}_n^1 - LATE^1) + \sqrt{n} \cdot (LATE^1 - LATE)$$

$$\xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^{2SLS^1}) + B$$

where B is defined as: $\sqrt{n} \cdot \left(\text{LATE}^1 - \text{LATE}\right) \xrightarrow[n\to\infty]{} B$

One can show that (Proposition 2 in WP):

$$\sqrt{n} \cdot \left(\text{LATE}^1 - \text{LATE}\right) = \frac{H \cdot P[G=0]}{P[D(1) > D(0)]} \cdot \left(\text{LATE}^1 - \text{LATE}^0\right)$$

$\left(\text{LATE}^1 - \text{LATE}^0\right)$ unrestricted $\Rightarrow$ asymptotic bias arbitrarily large.

Restricting treatment effect heterogeneity

Assumption 4 (vanishing treatment effect heterogeneity)

$$\left(\text{LATE}^1 - \text{LATE}^0\right) = O\left(\tfrac{1}{\sqrt{n}}\right) \Leftrightarrow \sup_n \left\{\sqrt{n} \cdot \left(\text{LATE}^1 - \text{LATE}^0\right)\right\} < \infty.$$

Why could this make sense?

→ Equivalently: "TE heterogeneity is of the order of sampling var."

→ Often times, applied researchers have a hard time detecting such heterogeneity ⇒ must be of this order in a lot of relevant applications

Restricting treatment effect heterogeneity

Assumption 4 (vanishing treatment effect heterogeneity)

$$\left(\text{LATE}^1 - \text{LATE}^0\right) = O\left(\tfrac{1}{\sqrt{n}}\right) \Leftrightarrow \sup_n \left\{\sqrt{n} \cdot \left(\text{LATE}^1 - \text{LATE}^0\right)\right\} < \infty.$$

Why could this make sense?

→ Equivalently: "TE heterogeneity is of the order of sampling var."

→ Often times, applied researchers have a hard time detecting such heterogeneity ⇒ must be of this order in a lot of relevant applications

Recap: credibility of our assumption

− in overpowered sciences (TE heterogeneity detected systematically)

\+ in underpowered sciences (economics and social sciences seem to fall in this category) where TE heterogeneity detected with difficulty

▸ Comparison with alternatives in recent lit. Conclusion

Roadmap

Framework

Proposed estimator

Asymptotic results under “standard” sequences

Asymptotic results under “local-to-zero” sequences

Finite-sample properties: Monte-Carlo simulations

Practical guidance and Conclusion

DGP 1 – a "best-case" scenario

DGP 1 in short:

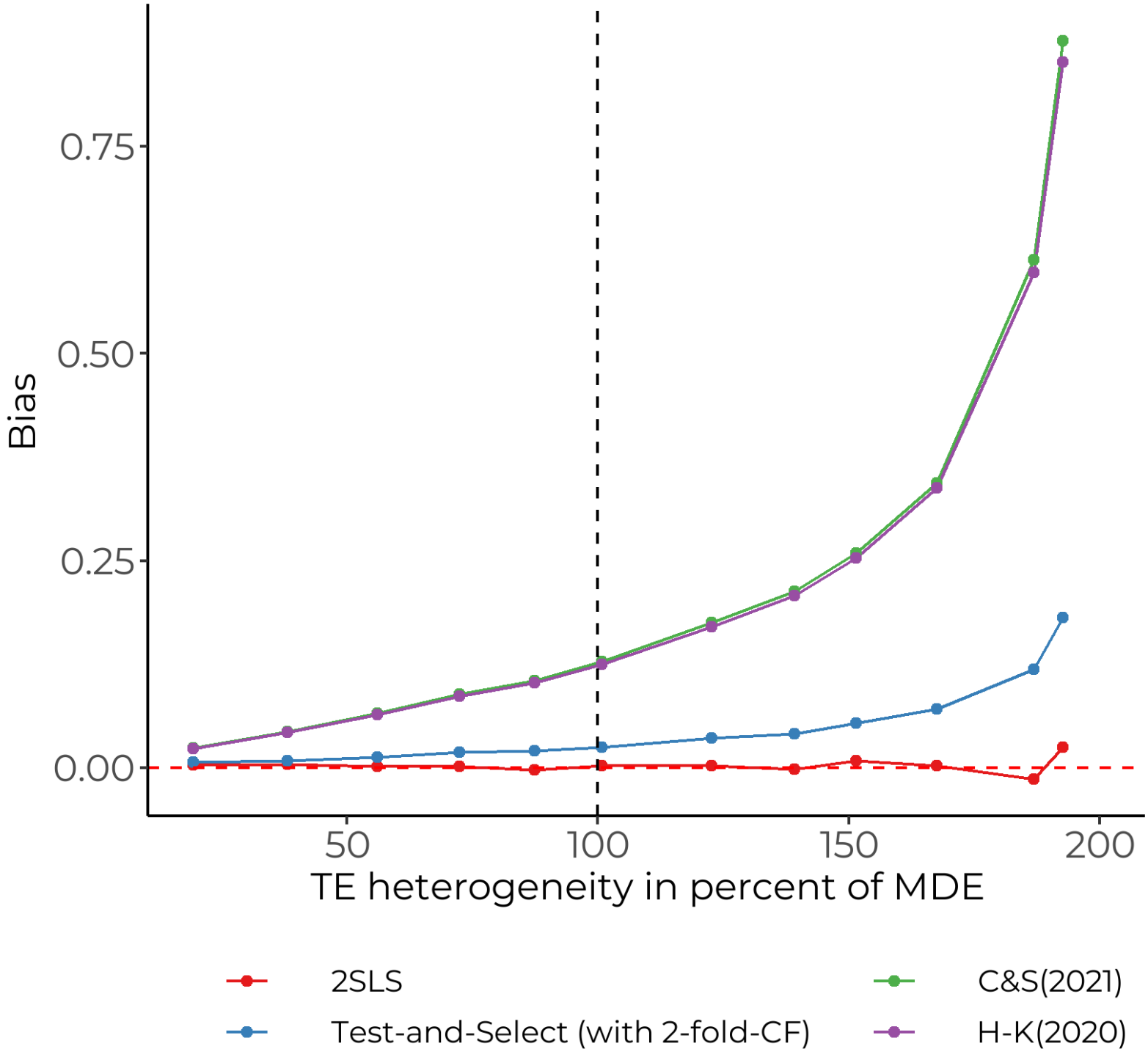
- N = 1000
- 10 groups with $\neq$ shares of compliers (overall $\pi = 0.25$)
  $\rightarrow$ some with $\pi^g = 0$ vs. others with above average π^g

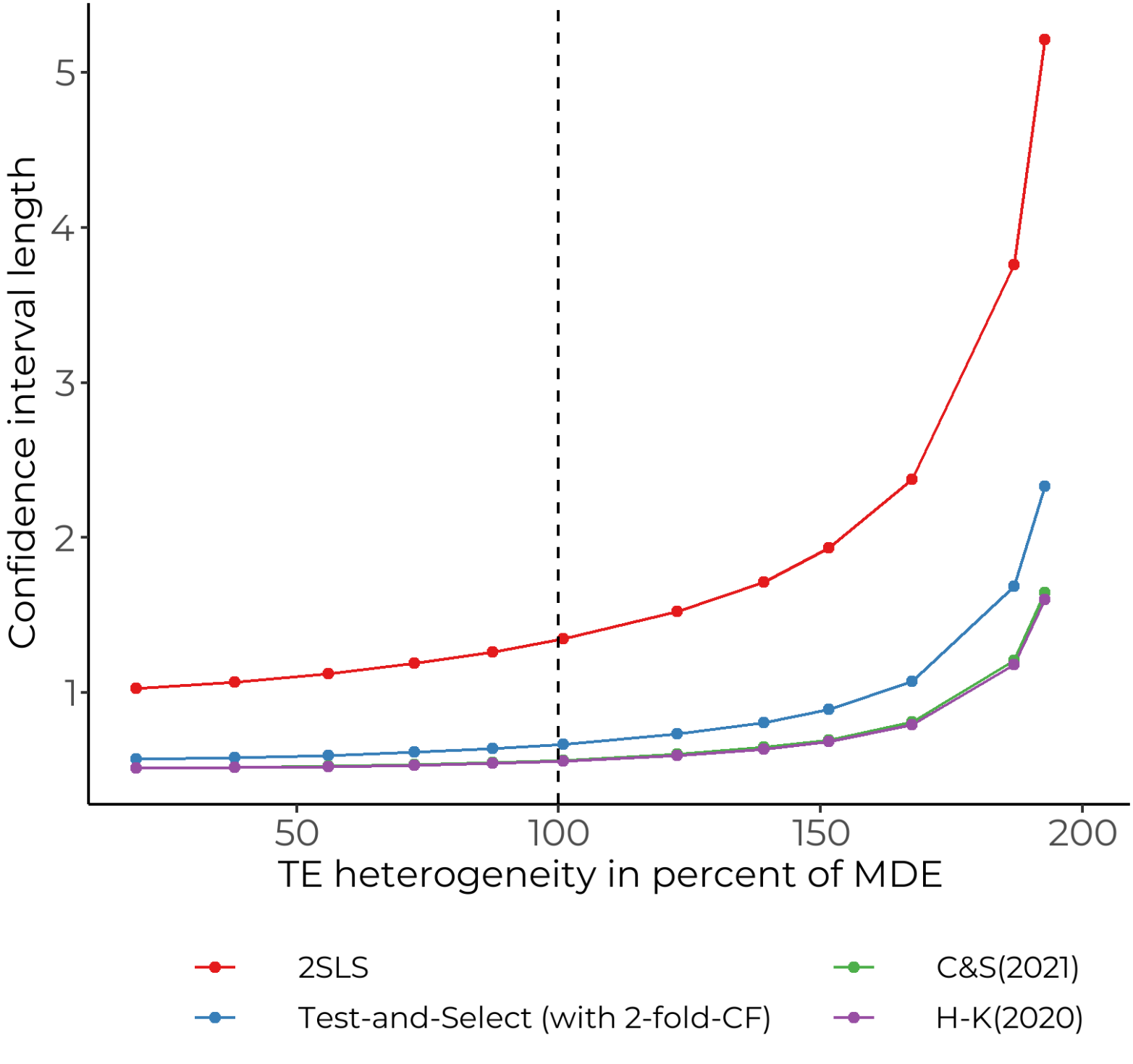
$$\pi_G = (\pi_1 = \pi_2 = \pi_3 = \pi_8 = \pi_9 = \pi_{10} = 0,$$
$$\pi_4 = \pi_7 \approx 0.25, \pi_5 = \pi_6 \approx 0.99)$$

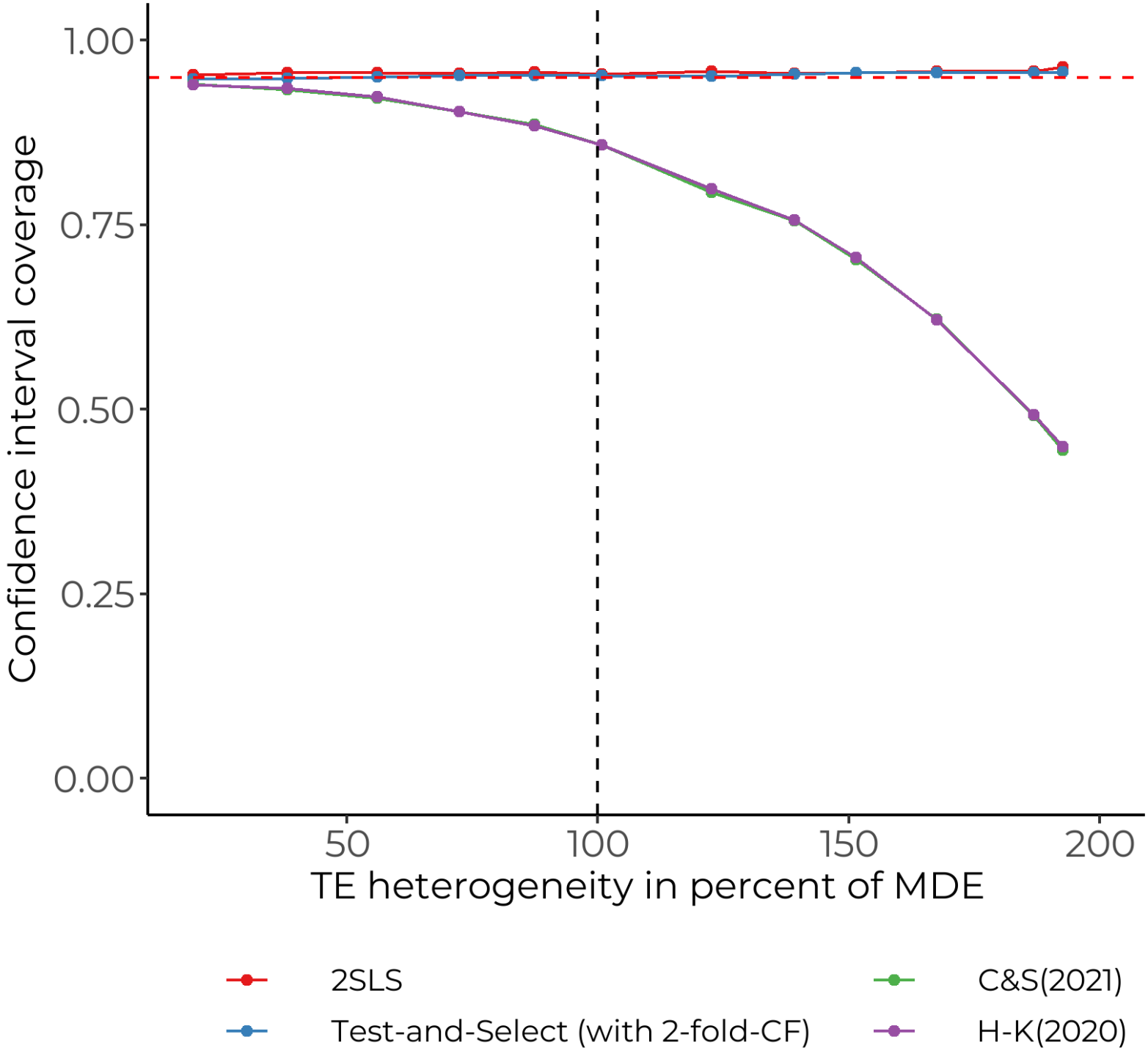
- $LATE^g \equiv E[Y(1) - Y(0) \mid D(1) > D(0), G = g]$ correlated with π^g

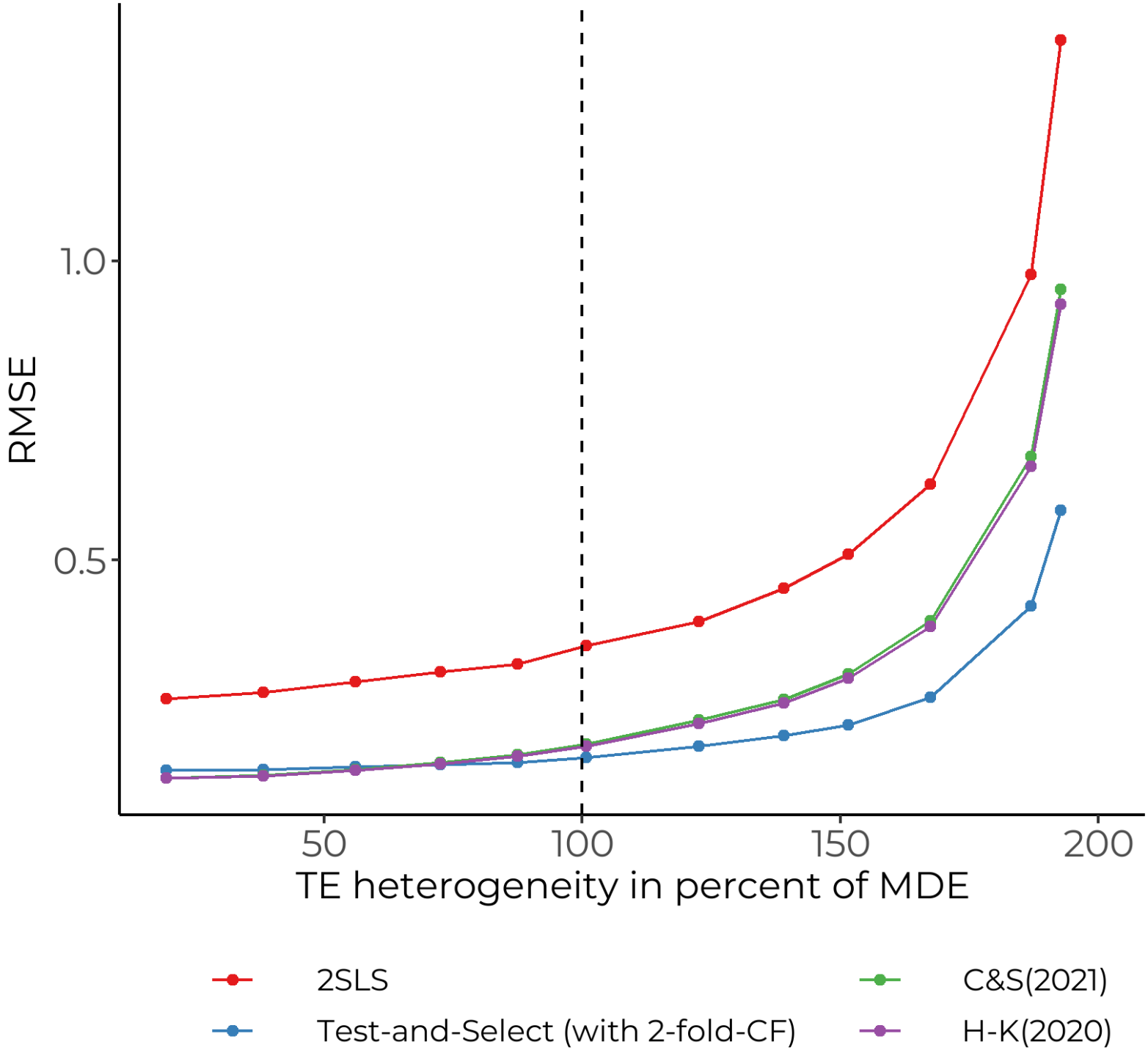
We vary the magnitude of the variance of treatment effects (TE), and scale it with respect to e^*, the minimum detectable effect (MDE).
$\rightarrow$ **Objective** = assess the robustness of our method (vs. alternatives) to various magnitudes of TE heterogeneity.

▸ Alternative estimators


0.75
0.50
0.25
0.00
Bias
50
100
150
200
TE heterogeneity in percent of MDE
2SLS
Test-and-Select (with 2-fold-CF)
C&S(2021)
H-K(2020)


Confidence interval length
1
2
3
4
5
50
100
150
200
TE heterogeneity in percent of MDE
2SLS
Test-and-Select (with 2-fold-CF)
C&S(2021)
H-K(2020)


1.00
0.75
0.50
0.25
0.00
Confidence interval coverage
50
100
150
200
TE heterogeneity in percent of MDE
2SLS
Test-and-Select (with 2-fold-CF)
C&S(2021)
H-K(2020)


RMSE
1.0
0.5
50
100
150
200
TE heterogeneity in percent of MDE
2SLS
Test-and-Select (with 2-fold-CF)
C&S(2021)
H-K(2020)

DGP 2 – introduction of "weak" first-stages

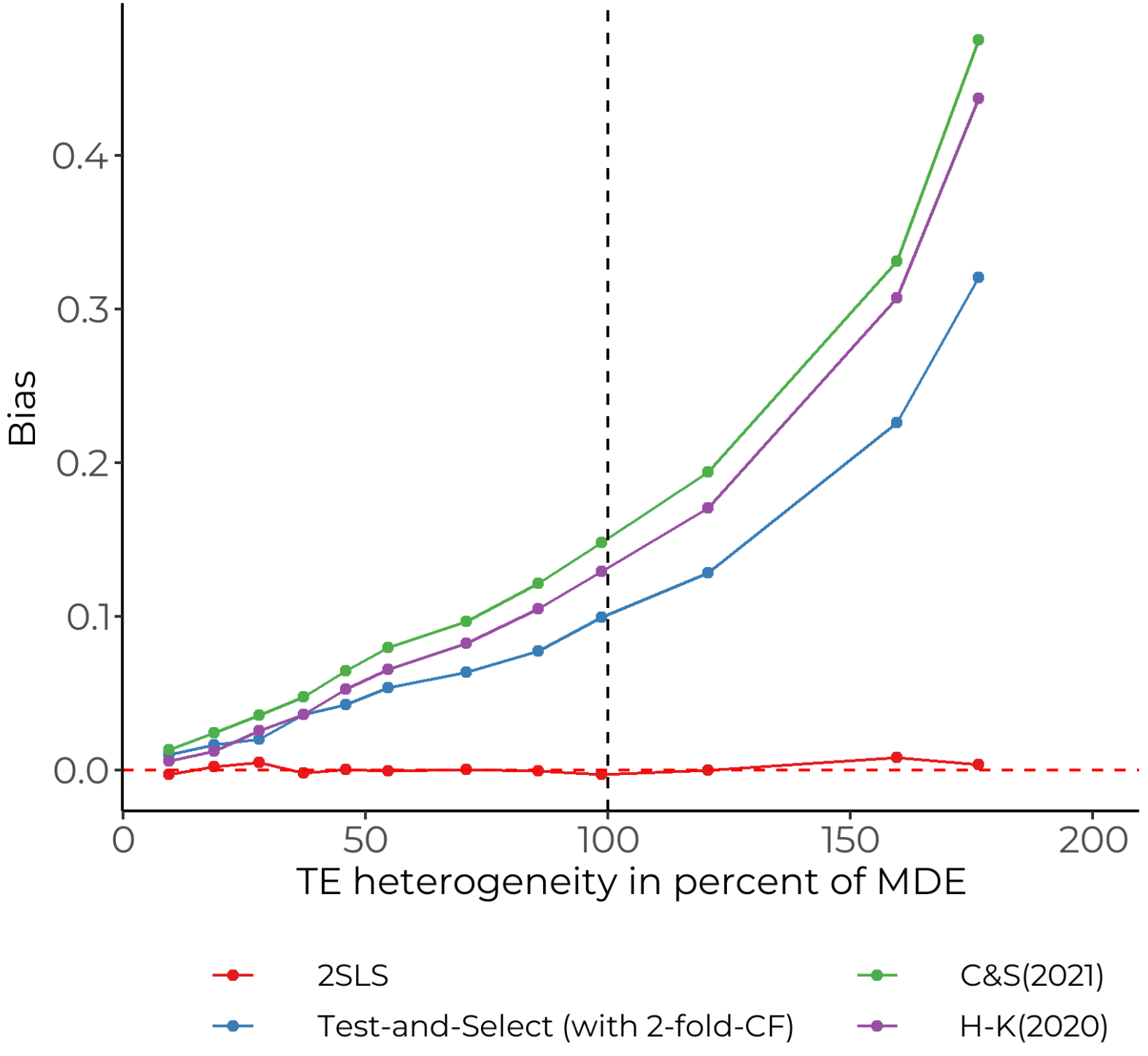
DGP 2 in short:

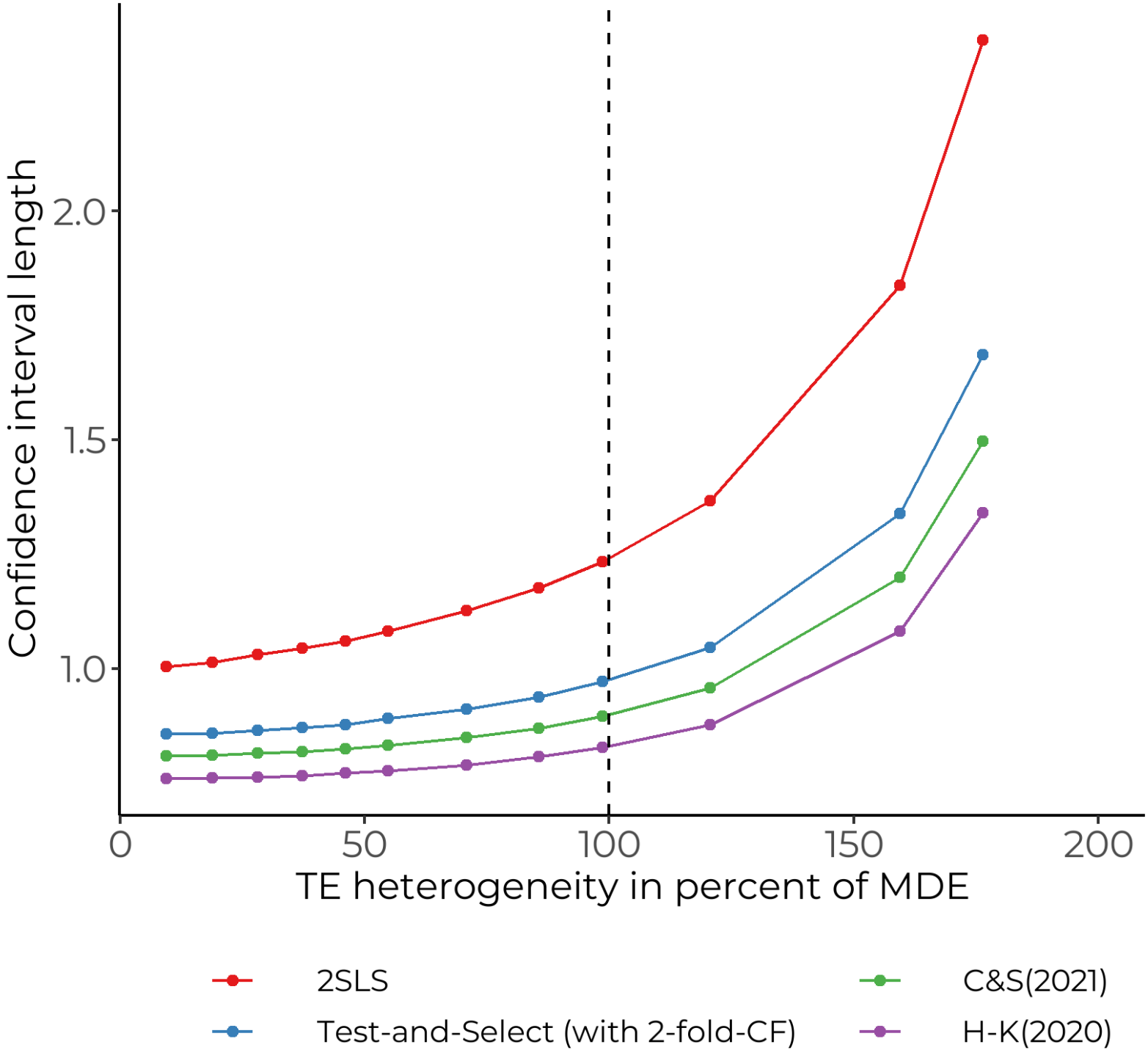
- N = 1000
- 10 groups with $\neq$ shares of compliers (overall $\pi = 0.25$)
  $\rightarrow$ some with π^g close to 0 vs. others with above average π^g

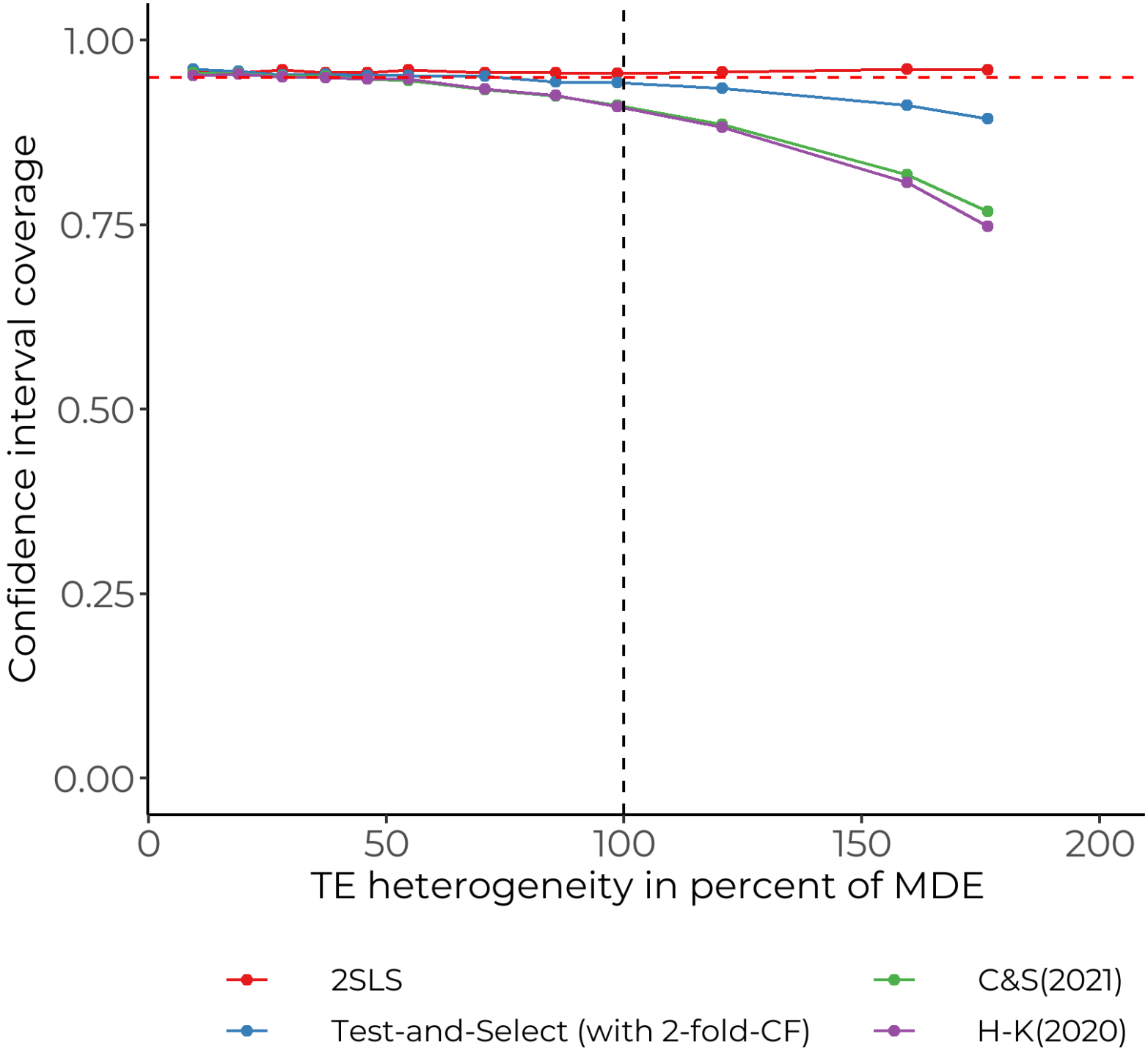
$$\pi_G = (\pi_1 = \pi_{10} \approx 0.001, \pi_2 = \pi_9 \approx 0.08,$$
$$\pi_3 = \pi_8 \approx 0.24, \pi_4 = \pi_7 \approx 0.40, \pi_5 = \pi_6 \approx 0.5)$$

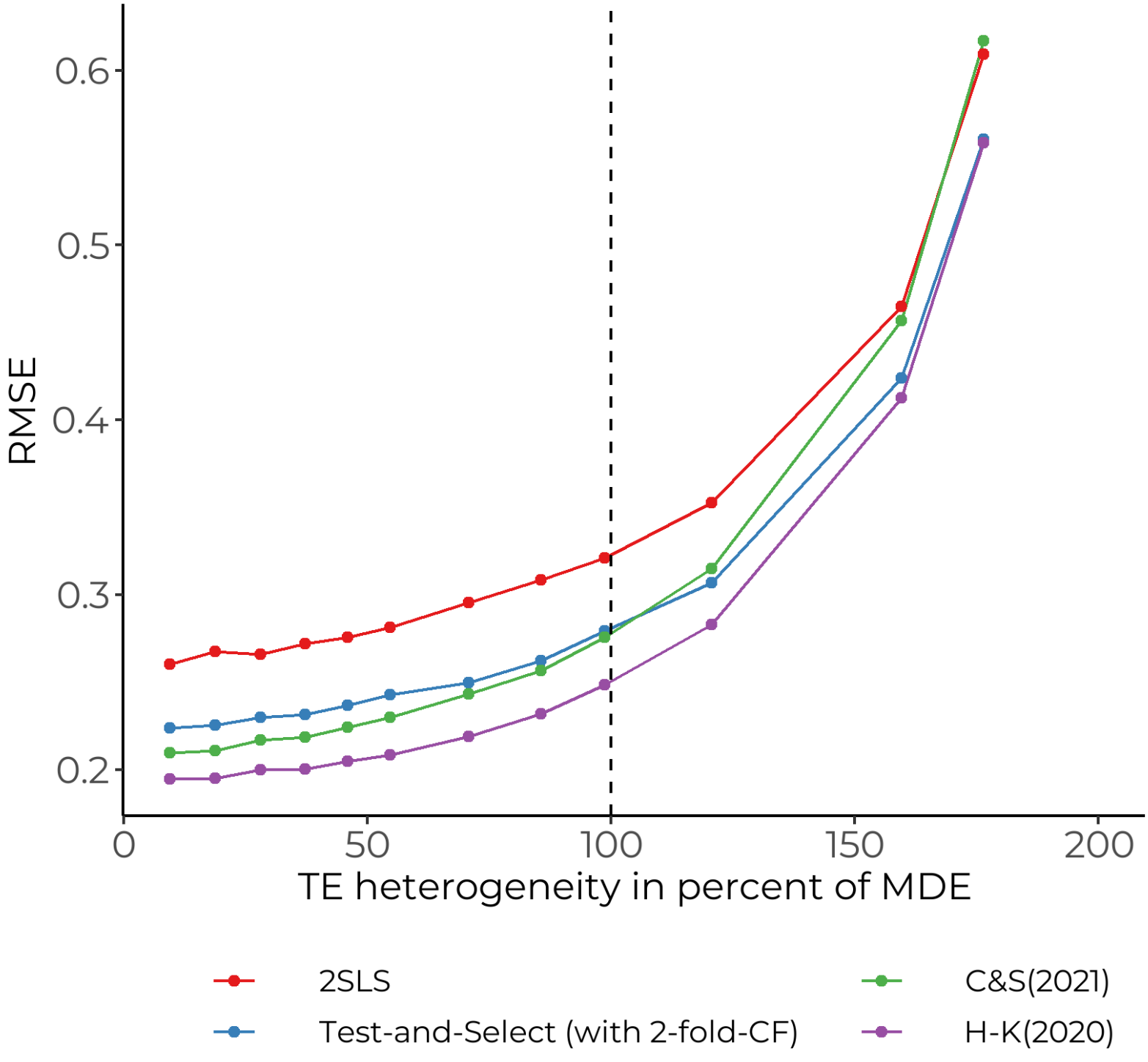
- $LATE^g \equiv E[Y(1) - Y(0) \mid D(1) > D(0), G = g]$ correlated with π^g

We vary the magnitude of the variance of treatment effects (TE), and scale it with respect to e^*, the minimum detectable effect (MDE).
$\rightarrow$ **Objective** = assess the robustness of our method (vs. alternatives) to various magnitudes of TE heterogeneity.


Bias
0.0
0.1
0.2
0.3
0.4
0
50
100
150
200
TE heterogeneity in percent of MDE
2SLS
Test-and-Select (with 2-fold-CF)
C&S(2021)
H-K(2020)


Confidence interval length
2.0
1.5
1.0
0
50
100
150
200
TE heterogeneity in percent of MDE
2SLS
Test-and-Select (with 2-fold-CF)
C&S(2021)
H-K(2020)


1.00
0.75
0.50
0.25
0.00
Confidence interval coverage
0
50
100
150
200
TE heterogeneity in percent of MDE
2SLS
Test-and-Select (with 2-fold-CF)
C&S(2021)
H-K(2020)


RMSE
0.6
0.5
0.4
0.3
0.2
0
50
100
150
200
TE heterogeneity in percent of MDE
2SLS
Test-and-Select (with 2-fold-CF)
C&S(2021)
H-K(2020)

Roadmap

Framework

Proposed estimator

Asymptotic results under “standard” sequences

Asymptotic results under “local-to-zero” sequences

Finite-sample properties: Monte-Carlo simulations

Practical guidance and Conclusion

Practical guidance: When should we use TS?

- ✓ When treatment effect heterogeneity fails to be detected
  - $\rightarrow$ relatively high p-value for test of H_0: "no TE heterogeneity"
  - $\rightarrow$ possibility to use ML-based tests for TE heterogeneity, e.g. in Chernozhukov et al., 2021 or Athey, Tibshirani, and Wager, 2019
- ✓ When some non-compliance can be predicted by covariates
  - $\rightarrow$ Use of contextual knowledge and/or ML prediction of TE het. in 1^{st} stage to use covariate info. in this step

Conclusion

- New estimator (TS) of the LATE exploiting 1^{st}-stage heterogeneity
  $\rightarrow$ builds on an intuitive estimation procedure
  $\rightarrow$ fixes pre-testing issue of naïve implementation
  $\rightarrow$ improves inference when non-compliance can be predicted

Conclusion

- New estimator (TS) of the LATE exploiting 1^{st}-stage heterogeneity
  $\rightarrow$ builds on an intuitive estimation procedure
  $\rightarrow$ fixes pre-testing issue of naïve implementation
  $\rightarrow$ improves inference when non-compliance can be predicted

- TS estimator remains 1^{st}-order unbiased for the LATE under reasonable assumptions on treatment effect heterogeneity
  $\rightarrow \neq$ from other estimators proposed in literature

Appendix

Variance gains

$$\sqrt{n}\left(\widehat{2SLS}_n - LATE\right) \to^d \mathcal{N}(0, V^{\text{2SLS}})$$

$$\sqrt{n}\left(\widehat{2SLS}_n^1 - \underbrace{LATE^1}_{=LATE}\right) \to^d \mathcal{N}(0, V^{\text{2SLS}^1})$$

A closer look:

$$V^{\text{2SLS}^1} = \frac{P[G=1]}{(\pi^1)^2} \cdot \left[\frac{1}{\mathrm{E}[Z]} \cdot \mathrm{Var}[\varepsilon|Z=1,G=1] + \frac{1}{1-\mathrm{E}[Z]} \cdot \mathrm{Var}[\varepsilon|Z=0,G=1]\right]$$

$$V^{\text{2SLS}} = \underbrace{\frac{1}{(\pi)^2}}_{=(\pi^1)^2} \cdot \left[\frac{1}{\mathrm{E}[Z]} \cdot \underbrace{V[\varepsilon|Z=1]}_{\geq P[G=1]\cdot \mathrm{Var}[\varepsilon|Z=1,G=1] \text{ (law of tot. var.)}} + \frac{1}{1-\mathrm{E}[Z]} \cdot \underbrace{V[\varepsilon|Z=0]}_{\geq P[G=1]\cdot \mathrm{Var}[\varepsilon|Z=0,G=1] \text{ (law of tot. var.)}}\right]$$

$$\geq V^{\text{2SLS}^1}$$

where $\varepsilon = Y - \mathrm{LATE} \cdot D$

Back

Restricting treatment effect heterogeneity

▸ Back ▸ Back MC

Comparison with alternatives in the literature

$$\left(\text{LATE}^1 - \text{LATE}^0\right) = O(1/\sqrt{n}) \quad \text{vs.} \quad \text{LATE}^1 = \text{LATE}^0$$

- TE homogenous: optimal IV is given by

$$\underbrace{\left(\text{E}[D \mid Z=1, X] - \text{E}[D \mid Z=0, X]\right)}_{=\pi^x} \cdot (Z - \text{E}[Z|X])$$

- TE heterogeneous: optimal IV targets a compliance-weighted LATE

$$\text{c. w. -LATE} \equiv \frac{\text{E}[\pi^X \cdot \text{LATE}^X \mid D(1) > D(0)]}{\text{E}[\pi^X \mid D(1) > D(0)]}$$

[Intuition: LATE^X reweighted by $(\pi^X)^2$ instead of π^X for the usual LATE.]

- TE heterogeneity = $O(1/\sqrt{n})$: 1st-order bias (Coussens and Spiess, '21)

$$\frac{\text{Cov}(\text{LATE}^X, \pi^X \mid D(1) > D(0))}{\text{E}[\pi^X \mid D(1) > D(0)]}$$

$\Rightarrow$ Usual CIs do not necessarily cover LATE at (1-nominal level)

Why should we care about LATE? ▸ Back

The choice of estimand

- When Z is the policy instrument at hand for the planner, the LATE (or equivalently the ITT) is a policy-relevant quantity.
- *But* "easiest" parameter to estimate is the compliance-weighted LATE

  [Coussens and Spiess (2021) target this estimand]

Statistical decision theory

- Null hypothesis testing (NHST) famously difficult to motivate...
- *Still*: motivation of NHST on ITT [conjecture: can be adapted for LATE] based on reference-dependence [Tetenov (2012), Banerjee et al. (2020)]
- Much more complicated for compliance-weighted LATE

  [Conjecture: most straightforward way = planner has a social welfare function with weights π^x instead of uniform weights. Why would that be?]

Why should we care about LATE? Back

What can we learn from compliance-weighted LATE?

- Inferring positivity of compliance-weighted LATE

  $\Rightarrow$ There is some groups w/ positive LATE...

  [Because compliance-weighted LATE is a convex avr. of conditional LATEs]

- If no TE heterogeneity:
  - compliance-weighted LATE = LATE = ATE

  $\Rightarrow$ optimal estimator = compliance-weighted estimator

- If TE heterogeneity: LATE can still be < 0
- A compliance-weighted estimator could dominate in RMSE...

  But no valid CIs can be derived if it is biased.

All in all: keeping the LATE as target parameter has important pros.