

# How Much Should We Trust Observational Estimates?

## Accumulating Evidence Using Randomised Controlled Trials with Imperfect Compliance

David Rhys Bernard   Gharad Bryan   Sylvain Chabé-Ferret   Jonathan de Quidt  
Jasmin Claire Fliegner   Roland Rathelot

PSE, LSE, TSE, QMUL, Manchester, IP Paris

March 2024

# Plan

Introduction

Setting

Sample and Context

Estimation

Results

Concluding Thoughts

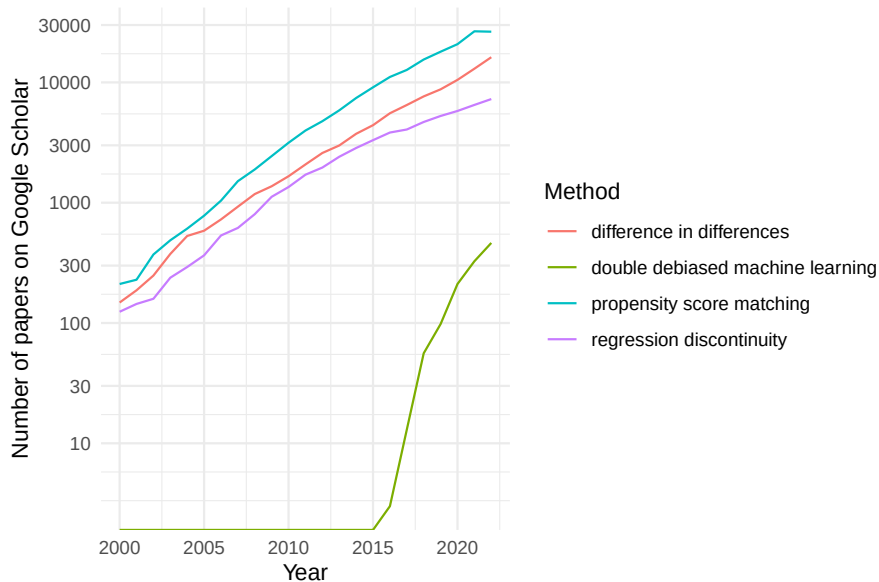
# So you want to evaluate this policy

- ▶ The Credibility Revolution
  - ▶ 1990s–2000s strong shift to more *credible* empirical methods in economics.
  - ▶ Emphasis on quasi-experimental methods and randomized experiments (IV, DiD, RD...)
- ▶ What about when there is no natural experiment?
  - ▶ Many important questions lack a credible natural experiment
  - ▶ Common response: give up and find another question
  - ▶ But policy makers, firms, NGOs don't have that option.



Source: Wikipedia

## Observational methods are very much in use



# Observational methods

- ▶ In the absence of a natural experiment we need to use “observational” alternatives (regression, matching, ML-based methods)
- ▶ What we hope for is **Conditional Independence**

$$(Y_i^1, Y_i^0) \perp D_i | X_i \quad (\text{CIA})$$

- ▶ Problem
  - ▶ In general, impossible to know if our observational method is performing well
  - ▶ Failure of the Conditional Independence Assumption (often called Selection on Unobservables) can give rise to a bias of **unknown direction and magnitude**

## What we do: treat observational bias as an empirical object

- ▶ Estimate the distribution of bias ( $B$ ): direction, size, **uncertainty**
  - ▶ “Honest” observational confidence intervals
- ▶ Method:
  - ▶ Collect a large data set of imperfect compliance RCTs (ICRCTs)
  - ▶ Within each RCT  $j$ :
    - ▶ obtain an experimental estimate  $EXP$  of the treatment effect on the treated
    - ▶ obtain an observational one  $OBS$  of the treatment effect on the treated
    - ▶ compute the  $B_j$  bias between the two
  - ▶ Combine these  $\{B_j\}_j$  to learn about the distribution of the bias

# Literature and contributions

## ► Literature

- LaLonde (1986)
- LaLonde's offspring: Heckman, Ichimura, Smith and Todd (1998), Dehejia and Wahba (1999, 2002), Smith and Todd (2005), Griffen and Todd (2017)
- Multi-study LaLonde-like analyses: Glazerman, Levy, Myers (2003), Gordon, Zettelmeyer, Bhargava, Chapsky (2019)

## ► Contributions

- Data standardisation: observational and experimental estimates from the same data
- Analysis standardisation: hands-off ML estimators
- Large set of studies, and outcomes measures: 44 studies
- Focus on uncertainty, rather than average bias

# Plan

Introduction

Setting

Sample and Context

Estimation

Results

Concluding Thoughts

## What we estimate, and how we adjust CIs

Policy maker estimates TOT with an observational method (regression, matching, ML)

She wishes:

$$\widehat{OBS}_p \sim \mathcal{N}(TOT_p, \sigma_\epsilon^2)$$

## What we estimate, and how we adjust CIs

Policy maker estimates TOT with an observational method (regression, matching, ML)

Reality:

$$\begin{aligned}\widehat{OBS}_p &\sim \mathcal{N}(TOT_p + B_p, \sigma_\epsilon^2) \\ B_p &\sim \mathcal{N}(\mu, \tau^2)\end{aligned}$$

## What we estimate, and how we adjust CIs

Policy maker estimates TOT with an observational method (regression, matching, ML)

Reality:

$$\begin{aligned}\widehat{OBS}_p &\sim \mathcal{N}(TOT_p + B_p, \sigma_\epsilon^2) \\ B_p &\sim \mathcal{N}(\mu, \tau^2)\end{aligned}$$

► If we could estimate

$$\hat{\mu}, \hat{\tau}^2, \hat{\sigma}_\mu^2$$

## What we estimate, and how we adjust CIs

Policy maker estimates TOT with an observational method (regression, matching, ML)

Reality:

$$\begin{aligned}\widehat{OBS}_p &\sim \mathcal{N}(TOT_p + B_p, \sigma_\epsilon^2) \\ B_p &\sim \mathcal{N}(\mu, \tau^2)\end{aligned}$$

- If we could estimate

$$\hat{\mu}, \hat{\tau}^2, \hat{\sigma}_\mu^2$$

- Appropriate adjusted two-sided  $100 * \delta$  % confidence interval for  $TOT$  is

$$\widehat{OBS}_p - \hat{\mu} \pm \Phi^{-1}\left(\frac{1+\delta}{2}\right) \underbrace{\sqrt{\hat{\sigma}_\epsilon^2 + \hat{\tau}^2 + \hat{\sigma}_\mu^2}}_{\text{effective SE}}$$

## How we estimate these quantities

- ▶ ICRCT: an RCT with imperfect compliance (ICRCT)
  - ▶ some assigned to  $T$  **do not** take up or
  - ▶ some assigned to  $C$  **do** take up
- ▶ Each ICRCT contains:
  - ▶ an experimental (EXP) study (by design)
  - ▶ an observational (OBS) study (comparing treated vs. untreated observations within randomly assigned groups)
- ▶ Comparing OBS to EXP gives measure of bias
  - ▶ Comparisons are **within study**, holding fixed: setting, program, outcome, measurement...
- ▶ 20 years of RCTs make our study feasible
  - ▶ We re-analyze 44 ICRCTs from IPA and JPAL

## How we estimate these quantities (2)

- ▶ For each study  $s \in S_{ICRCT}$ 
  1. Compute  $\widehat{EXP}_s$ : an experimental estimate for study  $s$
  2. Compute  $\widehat{OBS}_s$ : an observational estimate for study  $s$
  3. Estimate  $\widehat{B}_s = \widehat{OBS}_s - \widehat{EXP}_s$  (and  $\hat{\sigma}_{B_s}$ )
    - ▶ **Re-weight**: so  $EXP$  and  $OBS$  apply to the same population
    - ▶ **Hands-off methods**: remove researcher discretion
- ▶ Using estimates from all studies, random-effect meta-analysis recovers  $\{\hat{\mu}, \hat{\tau}^2, \hat{\sigma}_{\mu}^2\}$ :

$$\widehat{B}_s = \mu + \eta_s + \nu_s$$

$$\eta_s \sim N(0, \tau^2)$$

$$\nu_s \sim N(0, \hat{\sigma}_{B_s})$$

## Key assumption? A lack of knowledge

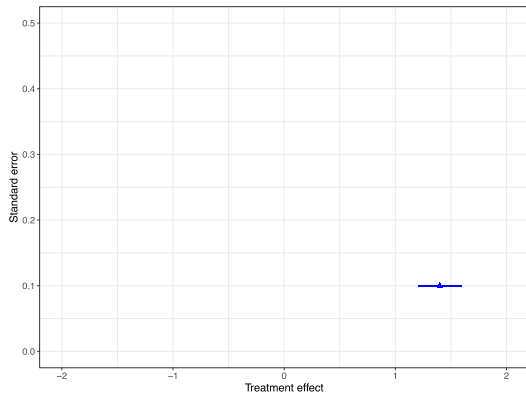
Key assumption for meta-analysis:

- ▶ Our ICRCTs are **exchangeable** with policy-makers' study of interest (and with each other)
- ▶ Practically  $B_s \sim N(\mu, \tau^2)$

For example:

- ▶ PM does not know that  $B_p$  has a lower mean than  $B_s$  for  $s \in S$
- ▶ PM does not know that  $\tau_p < \tau_s$  for  $s \in S$

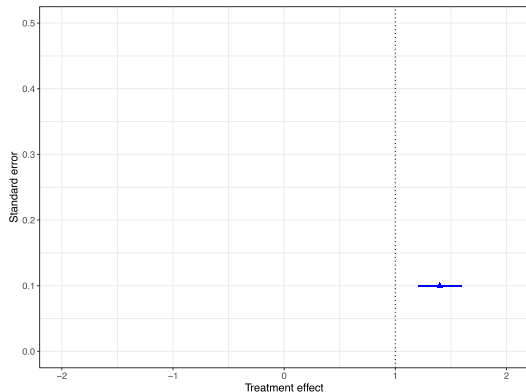
# Adjusting confidence intervals



► Large & significant  $\widehat{OBS}$

$$\widehat{OBS} \pm \Phi^{-1} \left( \frac{1 + \delta}{2} \right) \underbrace{\sqrt{\hat{\sigma}_\epsilon^2}}_{\text{effective SE}}$$

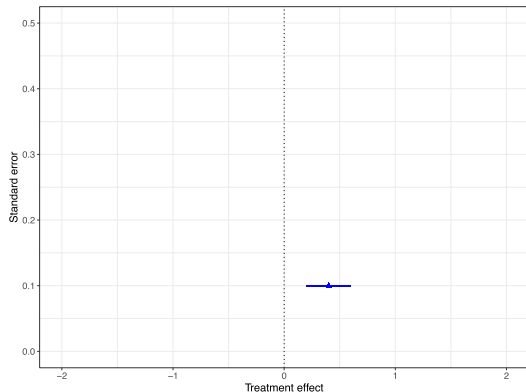
# Adjusting confidence intervals



- ▶ Large & significant  $\widehat{OBS}$
- ▶ Mean bias  $\hat{\mu} = 1$

$$\widehat{OBS} \pm \Phi^{-1} \left( \frac{1 + \delta}{2} \right) \underbrace{\sqrt{\hat{\sigma}_{\epsilon}^2}}_{\text{effective SE}}$$

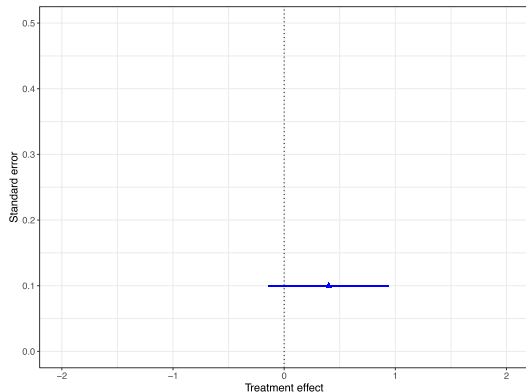
# Adjusting confidence intervals



- ▶ Large & significant  $\widehat{OBS}$
- ▶ Mean bias  $\hat{\mu} = 1$
- ▶ Subtract the mean bias

$$\widehat{OBS} - \hat{\mu} \pm \Phi^{-1} \left( \frac{1 + \delta}{2} \right) \underbrace{\sqrt{\hat{\sigma}_\epsilon^2}}_{\text{effective SE}}$$

# Adjusting confidence intervals



- ▶ Large & significant  $\widehat{OBS}$
- ▶ Mean bias  $\hat{\mu} = 1$
- ▶ Subtract the mean bias
- ▶ Correct standard error

$$\widehat{OBS} - \hat{\mu} \pm \Phi^{-1} \left( \frac{1 + \delta}{2} \right) \underbrace{\sqrt{\hat{\sigma}_\epsilon^2 + \hat{\tau}^2 + \hat{\sigma}_\mu^2}}_{\text{effective SE}}$$

# Plan

Introduction

Setting

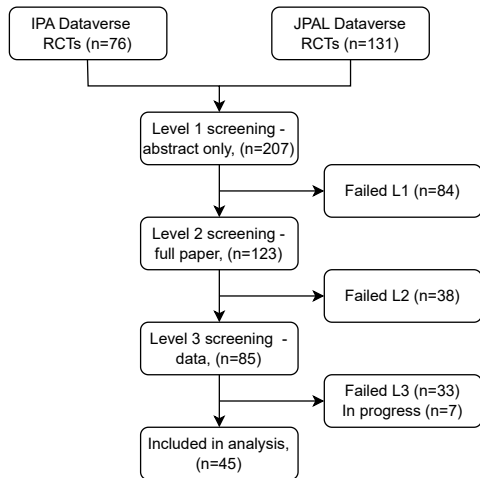
Sample and Context

Estimation

Results

Concluding Thoughts

# Data collection



## Systematic stepwise approach

- ▶ Systematic review of the J-PAL and IPA repositories on the Harvard Dataverse
  - ▶ Total of 207 complete data sets
- ▶ Screening process:
  - L1 Prescreening:** broad suitability
  - L2 In-depth screening:** eliminate papers that e.g. do not collect take-up or collect basic variables
  - L3 Data preparation:** information from the dataset and prepare a clean standardized format
- ▶ Identify as outcomes all LHS variables that appear in the paper/appendix
- ▶ “Primary” outcomes = mentioned in abstract
- ▶ 44 RCT datasets included in analysis

## Data: harmonization

We need all measurement on a common scale.

- ▶ Encode and align all outcomes by “social desirability” (= more is better, all else equal). Drop ambiguous cases.
- ▶ Divide by standard deviation in the control group (effect sizes)
- ▶ Identify “primary” outcomes (mentioned in abstract)
- ▶ Aggregate into indices using precision weights (for all outcomes, and for primary outcomes only).
  - ▶ If still multiple specifications per paper, keep the most powerful one.

Our preferred estimates use the aggregated primary outcomes: 1 per study.

# Outcome variables

## Normalization

- ▶ Direction: positive means socially desirable
- ▶ Units: standard deviation of the control

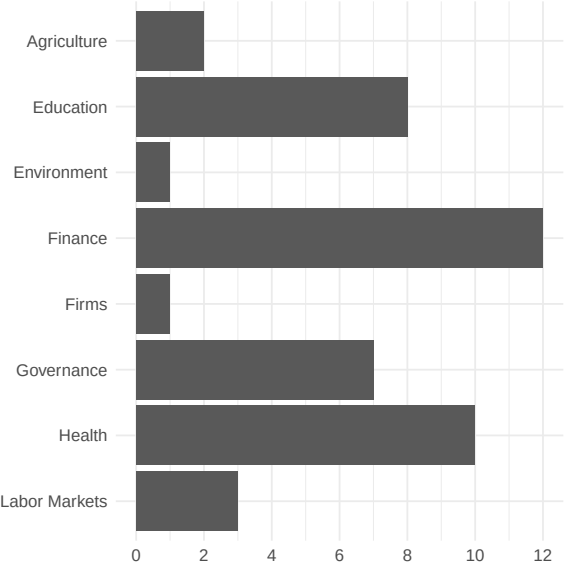
We take all outcomes in paper, and create four subsets

- ▶ Primary/All outcomes
- ▶ Aggregated/Disaggregated

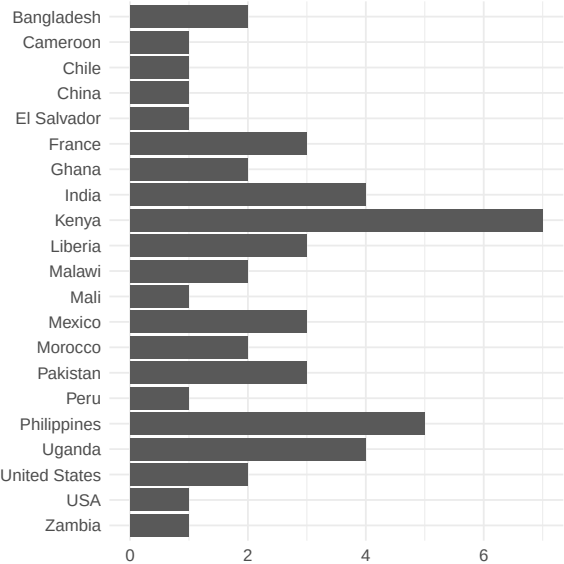
Two decisions: how to aggregate and what is primary?

- ▶ Primary outcomes are those mentioned in the abstract
- ▶ Aggregation takes precision-weighted mean (agnostic welfare measure)

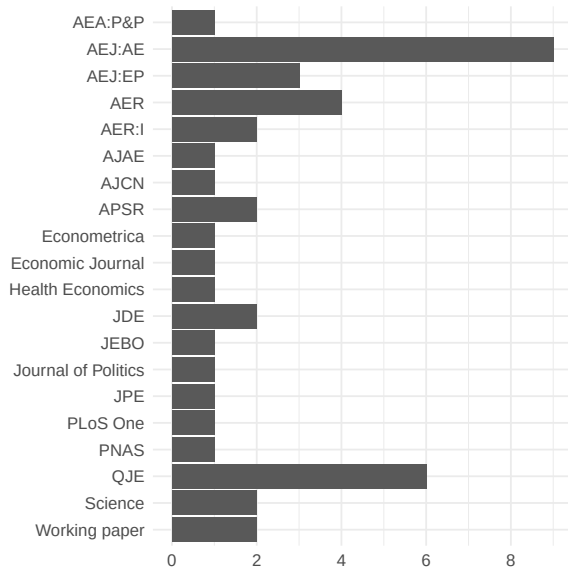
# Sample: Sector



# Sample: Country



## Sample: Journal



# Plan

Introduction

Setting

Sample and Context

Estimation

Results

Concluding Thoughts

## Estimating bias

Our starting point will be an imperfect-compliance RCT (ICRCT).

## Estimating bias

Our starting point will be an imperfect-compliance RCT (ICRCT).

- ▶  $R_i$ : randomized offer to take a program (treatment vs. control)
- ▶  $D_i$ : actual participation in the program (participants vs. non participants)

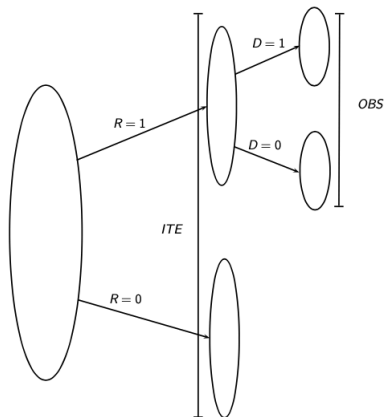
Imperfect compliance  $\Rightarrow D_i \neq R_i$  for some participants.

Two kinds of ICRCT

- ▶ Eligibility design: treatment group eligible for program, control group ineligible.
- ▶ Encouragement design: everyone eligible, treatment group **encouraged** to take up.

We'll always assume our ICRCTs satisfy Exogeneity, First stage, SUTVA, Exclusion, and Monotonicity

# Eligibility Designs



- ▶ Individuals allocated to the control group cannot participate: ( $Pr(D_i = 1 | R_i = 0) = 0$ )
- ▶ Experimental estimator

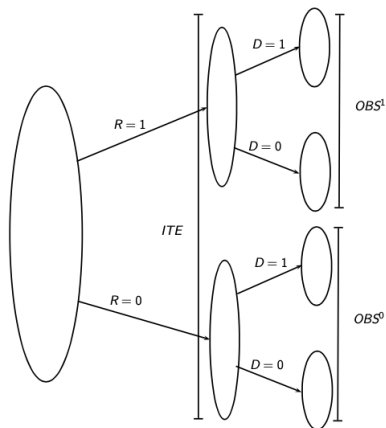
$$TOT^{EXP} = \frac{ITE}{Takeup} \rightarrow TOT$$

- ▶ Observational estimator compares treated eligible to non-treated eligible participants, conditional on observables.

$$TOT_X^{OBS} = \mathbb{E}[Y_i^1 | X_i, D_i = 1, R_i = 1] \\ - \mathbb{E}[Y_i^0 | X_i, D_i = 0, R_i = 1]$$

- ▶  $= TOT$  under **if CIA, SUTVA, common support hold.**

# Encouragement designs



- ▶ Experimental Wald estimator recovers Treatment Effect for Compliers (TOC):

$$TOC^{EXP} = \frac{ITE}{\text{Share of compliers}} \rightarrow TOC$$

Recall: compliers are those induced to take up by the encouragement

- ▶ Compute  $TOT^{OBS,r}$  for each treatment arm  $r$ .
- ▶ We can construct an observational estimate of TOC:

$$TOC_X^{OBS} = \frac{TOT_X^{OBS,1} \Pr(D_i = 1|R_i = 1) - TOT_X^{OBS,0} \Pr(D_i = 1|R_i = 0)}{\Pr(D_i = 1|R_i = 1) - \Pr(D_i = 1|R_i = 0)}$$

## Estimating $B$

For study  $s$  and outcome  $Y$

1. Compute experimental estimates for study  $s$
2. Compute observational estimates using a “best-available” observational method (e.g. Double-Debiased Machine Learning)
3. Then estimate  $B_s$  for that study

$$\hat{B}_s = \widehat{TOT}_s^{OBS} - \widehat{TOT}_s^{EXP} \quad \text{(Eligibility)}$$

$$\hat{B}_s = \widehat{TOC}_s^{OBS} - \widehat{TOC}_s^{EXP} \quad \text{(Encouragement)}$$

# Estimation

- ▶ Experimental: Instrument take-up with treatment, control for strata dummies.
- ▶ Observational:
  - ▶ Naive “With-Without” comparison (WW) (identifies the bias without any controls)
  - ▶ Post Double Selection Lasso (PDSL, Belloni et al., 2014)
  - ▶ Double-debiased Machine Learning (DDML, Chernozhukov et al., 2018)
- ▶ PDSL and DDML are “hands-off” observational approaches
  - ▶ PDSL: use LASSO to select  $X$  that are predictive of  $Y$  and  $D$ . Regress  $Y$  on  $D$  and selected covariates.
  - ▶ DDML: split sample, on training set flexibly predict  $Y$  and  $D$ , on holdout set perform “partially linear” regression that differences out those predictions. Repeat and take the average.
  - ▶ Both can handle large numbers of covariates (even  $k > N$ ).
  - ▶ No researcher input beyond setting tuning parameters (we use defaults)

## Aggregating estimates

- Suppose we have  $s = 1, \dots, S$  studies, with one estimated bias per study

$$\widehat{B}_s = \widehat{OBS}_s - \widehat{EXP}_s$$

- We assumed that  $B_s \sim \mathcal{N}(\mu, \tau^2)$  so

$$\widehat{B}_s = \mu + \eta_s + \nu_s \quad (\text{Mean bias})$$

$$\eta_s \sim \mathcal{N}(0, \tau^2) \quad (\text{Fundamental variance})$$

$$\nu_s \sim \mathcal{N}(0, \hat{\sigma}_s) \quad (\text{Sampling error})$$

- We use random-effects meta-analysis (weights given by  $\hat{\sigma}_s$ ) to recover  $\hat{\mu}, \hat{\sigma}_\mu^2, \hat{\tau}^2$
- If we have multiple outcomes per study, allow for within-study correlation between biases using a multi-level model.

# Plan

Introduction

Setting

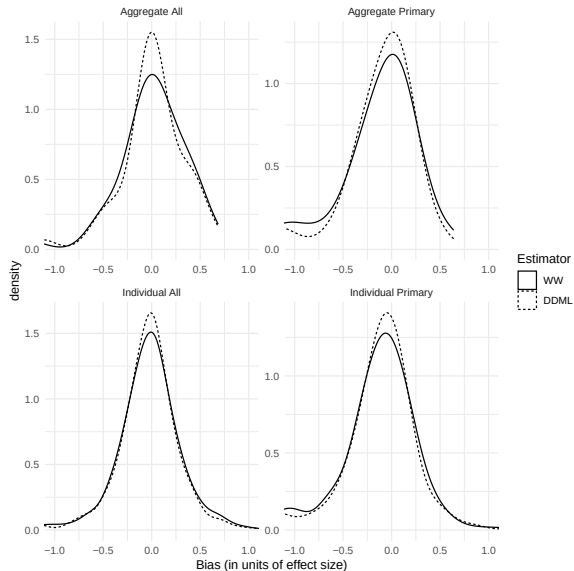
Sample and Context

Estimation

**Results**

Concluding Thoughts

# Distribution of selection bias



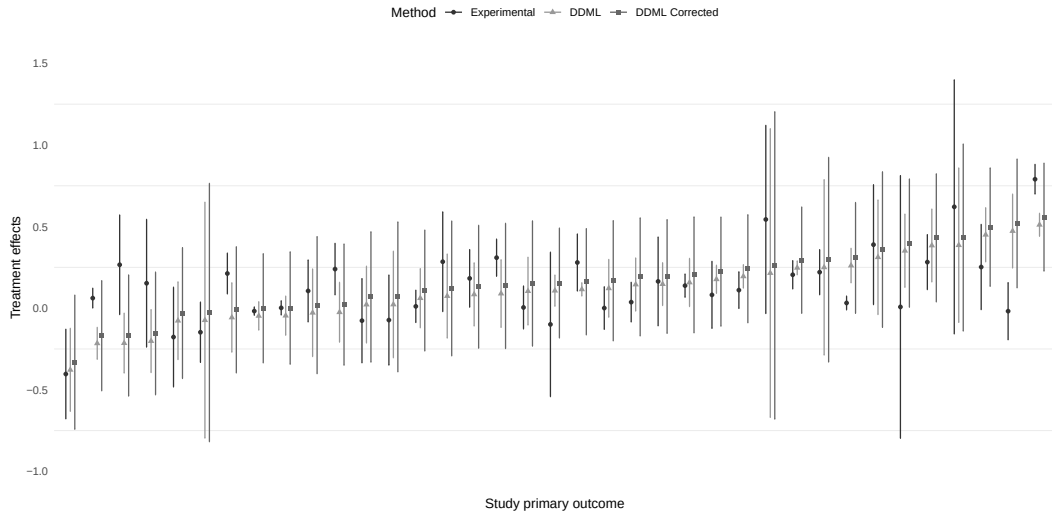
## 4 main results

1. Average bias very close to zero
2. Variance is large
3. Distribution not far from normal
4. Conditioning on observed covariates decreases selection bias

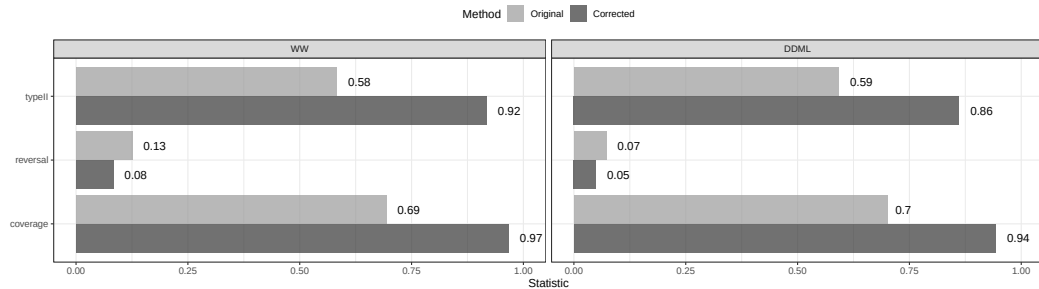
## Meta-analytic estimates

|                                             | TE<br>(1) | WW<br>(2) | DDML<br>(3) | PDSL<br>(4) |
|---------------------------------------------|-----------|-----------|-------------|-------------|
| <i>Panel A: Aggregated primary outcomes</i> |           |           |             |             |
| Mean ( $\hat{\mu}$ )                        | 0.171     | -0.046    | -0.053      | -0.047      |
| SE ( $\hat{\sigma}_{\mu}$ )                 | (0.042)   | (0.042)   | (0.037)     | (0.035)     |
| Standard deviation ( $\hat{\tau}$ )         |           | 0.201     | 0.165       | 0.161       |
| Effective.SE                                |           | 0.206     | 0.169       | 0.165       |
| Num.obs.                                    | 42        | 42        | 42          | 42          |
| <i>Panel B: Individual primary outcomes</i> |           |           |             |             |
| Mean ( $\hat{\mu}$ )                        | 0.126     | -0.052    | -0.074      | -0.052      |
| SE ( $\hat{\sigma}_{\mu}$ )                 | (0.031)   | (0.036)   | (0.031)     | (0.031)     |
| Standard deviation ( $\hat{\tau}$ )         |           | 0.231     | 0.199       | 0.199       |
| Effective.SE                                |           | 0.234     | 0.202       | 0.202       |
| Num.obs.                                    | 264       | 264       | 264         | 264         |

# Corrected confidence intervals for $\widehat{OBS}$

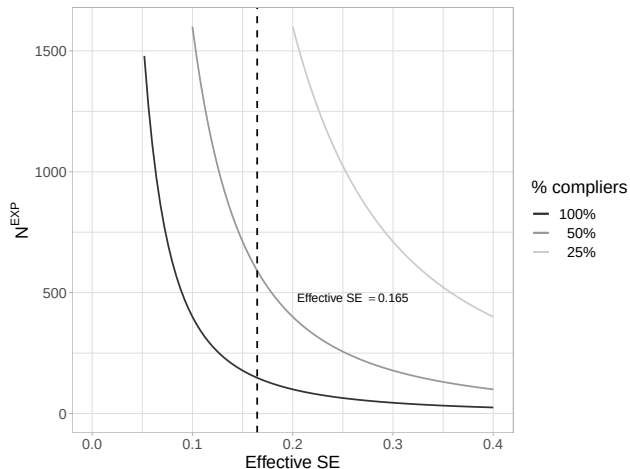


# Performance of our proposed correction

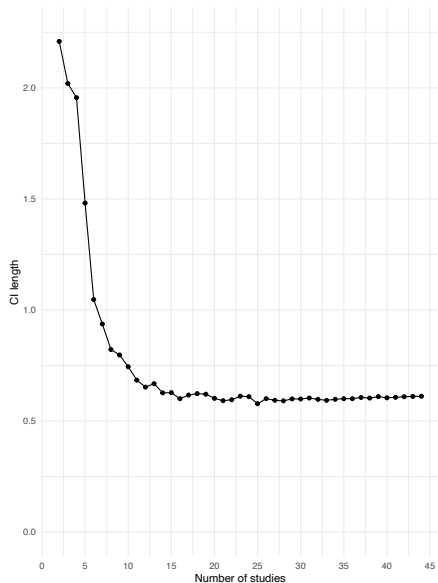


## Equivalent sample size

What is the minimum sample size of an RCT to reach the same precision as an observational study with infinite sample size?

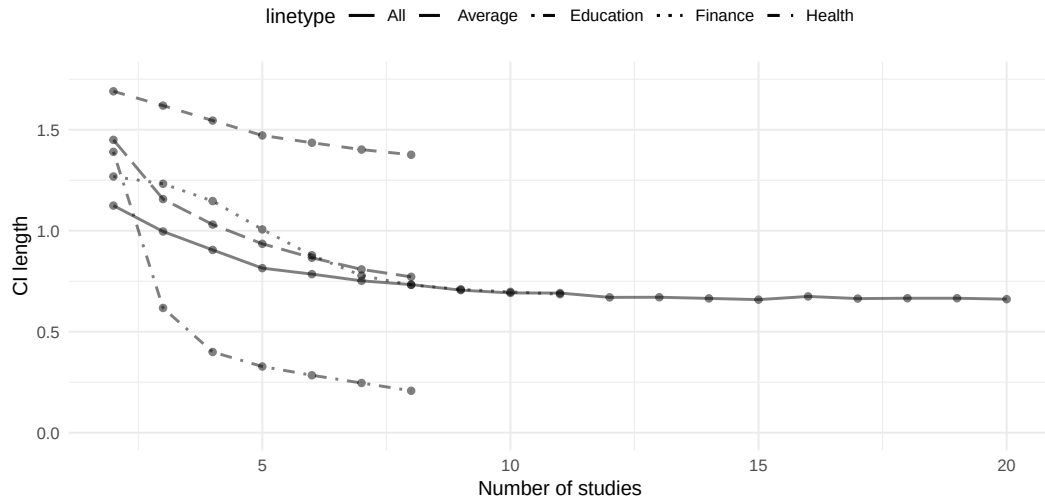


# The Value of additional ICRCTs



- ▶ One more ICRCT reduces the size of the confidence interval
- ▶ Most gains will come from allowing narrower meta-analysis focusing on 10–15 most-relevant studies.

# The Value of additional ICRICTs



# Plan

Introduction

Setting

Sample and Context

Estimation

Results

Concluding Thoughts

## Take-aways

1. Average bias close to zero:  $\mu \approx .03$  s.d.
  - ▶ Little systematic over/underestimation of welfare impacts

## Take-aways

1. Average bias close to zero:  $\mu \approx .03$  s.d.
  - ▶ Little systematic over/underestimation of welfare impacts
2. Large uncertainty due to selection bias:  $\hat{\tau}^2 + \hat{\sigma}_{\mu}^2 \approx 0.16$  s.d.
  - ▶ This is the effective SE when  $N^{OBS} \rightarrow \infty$

## Take-aways

1. Average bias close to zero:  $\mu \approx .03$  s.d.
  - ▶ Little systematic over/underestimation of welfare impacts
2. Large uncertainty due to selection bias:  $\hat{\tau}^2 + \hat{\sigma}_{\mu}^2 \approx 0.16$  s.d.
  - ▶ This is the effective SE when  $N^{OBS} \rightarrow \infty$
3. RCTs dominate  $N^{OBS} = \infty$  at low sample sizes
  - ▶ A perfect-compliance RCT has more power at  $N^{EXP} = 148$
  - ▶ A 25%-compliance RCT has more power at  $N^{EXP} = 2350$

# Take-aways

1. Average bias close to zero:  $\mu \approx .03$  s.d.
  - ▶ Little systematic over/underestimation of welfare impacts
2. Large uncertainty due to selection bias:  $\hat{\tau}^2 + \hat{\sigma}_{\mu}^2 \approx 0.16$  s.d.
  - ▶ This is the effective SE when  $N^{OBS} \rightarrow \infty$
3. RCTs dominate  $N^{OBS} = \infty$  at low sample sizes
  - ▶ A perfect-compliance RCT has more power at  $N^{EXP} = 148$
  - ▶ A 25%-compliance RCT has more power at  $N^{EXP} = 2350$
4. CI adjustment 'works'
  - ▶ 70% coverage before adjustment, 94% with adjusted CI
  - ▶ At the cost of losing power

## Take-aways

1. Average bias close to zero:  $\mu \approx .03$  s.d.
  - ▶ Little systematic over/underestimation of welfare impacts
2. Large uncertainty due to selection bias:  $\hat{\tau}^2 + \hat{\sigma}_{\mu}^2 \approx 0.16$  s.d.
  - ▶ This is the effective SE when  $N^{OBS} \rightarrow \infty$
3. RCTs dominate  $N^{OBS} = \infty$  at low sample sizes
  - ▶ A perfect-compliance RCT has more power at  $N^{EXP} = 148$
  - ▶ A 25%-compliance RCT has more power at  $N^{EXP} = 2350$
4. CI adjustment 'works'
  - ▶ 70% coverage before adjustment, 94% with adjusted CI
  - ▶ At the cost of losing power
5. Methods differ in performance
  - ▶ DDML generally best

## Next steps

- ▶ We could reduce the uncertainty further
  - ▶ Switch from exchangeability to *conditional* exchangeability
    - ▶ Choosing how to condition in a clever manner
  - ▶ Run more ICRCTs (joint effort)
    - ▶ Particularly for areas with few existing studies.

Perhaps the future is observational studies with slightly larger standard errors!