Lurking inferential monsters?

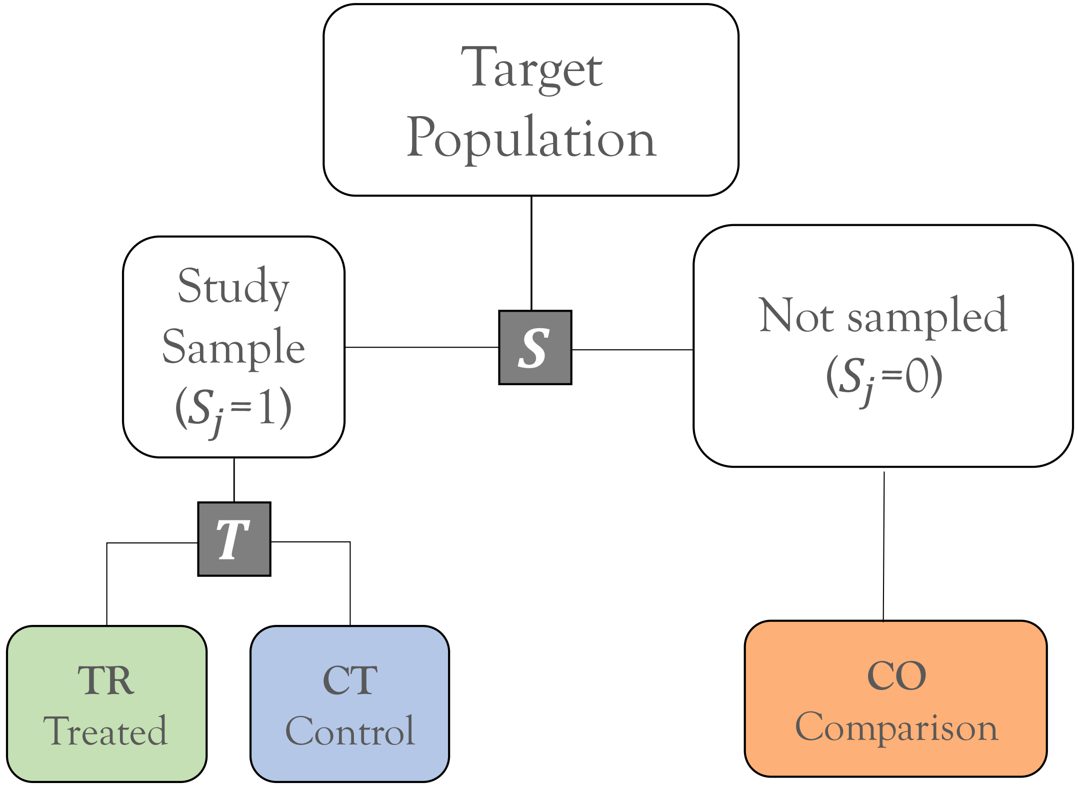
Quantifying Selection Bias in Evaluations of School Programs

Ben Weidmann
Harvard Graduate School of Education
& Education Endowment Foundation

SREE
Sep 2020

Weidmann, Ben, and Luke Miratrix. "Lurking Inferential Monsters: Quantifying Selection Bias in Evaluations of School Programs." *Journal of Policy Analysis and Management*.

Within-study comparisons



$$\hat{\tau}_{RCT} = \bar{Y}_{TR} - \bar{Y}_{CT} \qquad \hat{\tau}_{Non-RCT} = \bar{Y}_{TR} - \bar{Y}_{CO}$$

$$\hat{\beta} = \hat{\tau}_{Non-RCT} - \hat{\tau}_{RCT}$$

$$\hat{\beta} = \bar{Y}_{CT} - \bar{Y}_{CO}$$

Sources: Imai, King and Stuart (2008); Wong and Steiner (2016); LaLonde (1986)

Data

Sources	Data
Archive of experiments (RCTs)	Information on which schools/pupils participated in experiments
Pupil	
• Performance tables	Academic attainment in maths and English (outcomes=grade 6; pre-test = grade 2)
• Pupil Census	Demographics (age, rurality, gender)
School	
• Performance tables	Average attainment (level and change over time)
• School census	School size & type (academy status, # of pupils)
• School workforce	Staffing (e.g. teacher:pupil ratio)
• School finance	Budget (£/pupil; spending on 'outside services')
• Ofsted	Most recent Ofsted evaluation [Ofsted = official inspection body]
Neighbourhood	
• Indices of deprivation	Children Deprivation Index (IDACI), crime, housing

Interventions (1 of 2)

Intervention	ID	Brief description of intervention	n_schools* (n_pupils)	Reference
Affordable Online Maths Tuition	am	1-on-1 online tutoring, for grade 6's by math graduates in India and Sri Lanka. ~45 mins per week for 27 weeks.	64 (3,106)	Torgerson et al. (2016)
Changing Mindsets	cm	Professional development course for primary school teachers in how to develop Growth Mindset in pupils.	30 (1,505)	Rienzo et al. (2015)
Chess in Schools	chs	Grade 5 students taught chess by experienced chess tutor, instead of music or PE, over 30 weeks.	100 (4,009)	Jerrim et al. (2016)
Dialogic Teaching	dt	Grade 5 teachers trained in techniques to encourage dialogue, argument and oral explanation during class time	78 (4,958)	Jay et al. (2017)
Flipped Learning	Fl	Grade 5 pupils learn core math content online, outside of class time. Classes were used to reinforce/clarify ideas.	24 (1,214)	Rudd, Aguilera, Elliot, and Chambers (2017)
Hampshire Hundreds	hh	Professional development for primary schools teachers in strategies to reduce educational achievement gaps.	36 (2,048)	McNally et al. (2014)
Learner Response System	lrs	Handheld devices used in grades 5 and 6, to provide teachers with real-time information about pupil knowledge	97 (3,213)	Wiggins, Sawtell, and Jerrim (2017)

Intervention	ID	Brief description of intervention	n_schools* (n_pupils)	Reference
Magic Breakfast	mb	Providing nutritious breakfast to primary school students for most of the 2014-15 academic year.	98 (4,038)	Crawford et al. (2016)
Mind the Gap	mtg	Teacher training and parent workshops (over a 5 week period) to help grade 4 students be more 'meta-cognitive'.	45** (1,603)	Dorsett et al. (2014)
Philosophy for Children	p4c	Dialogic teaching of philosophical issues to children in grades 4 and 5, over a period of 11 months.	48 (1,529)	Gorard et al. (2015)
ReflectEd	ref	Weekly lessons where grade 5's learn strategies to monitor/manage their own learning (over 6 months)	33 (1,858)	Motteram et al. (2016)
Shared Maths	sm	Cross-age peer math tutoring: older pupils (grade 6) work with younger ones (grade 4) for 20 mins per week for 2 years.	82 (3,167)	Lloyd et al. (2015)
Talk of the Town	tott	Whole-school intervention to help support the development of children's speech, language and communication.	64 (3,299)	Thurston et al. (2016)
Thinking, Doing, Talking Science	ttds	5 day's professional development for grade 5 teachers, with the aim of making science more practical and engaging.	42 (1,513)	Hanley et al. (2015)

*n_schools (pupils) describes the number of schools and pupils included in the original RCT evaluations at randomization.
**Figures based on the EEF Archive, rather than the published report, as the latter did not include the number of students at randomization.

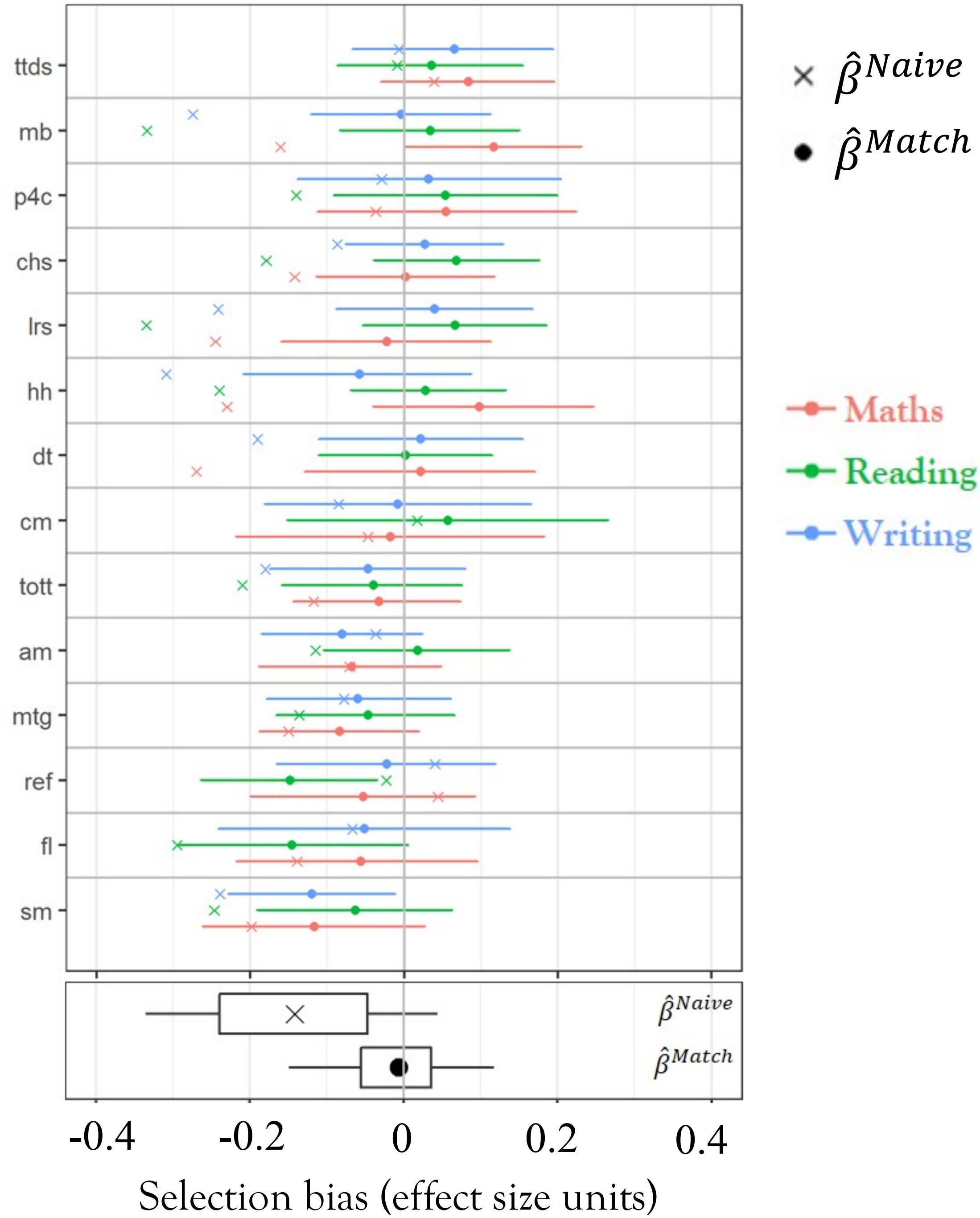
Estimating selection bias

Naïve bias

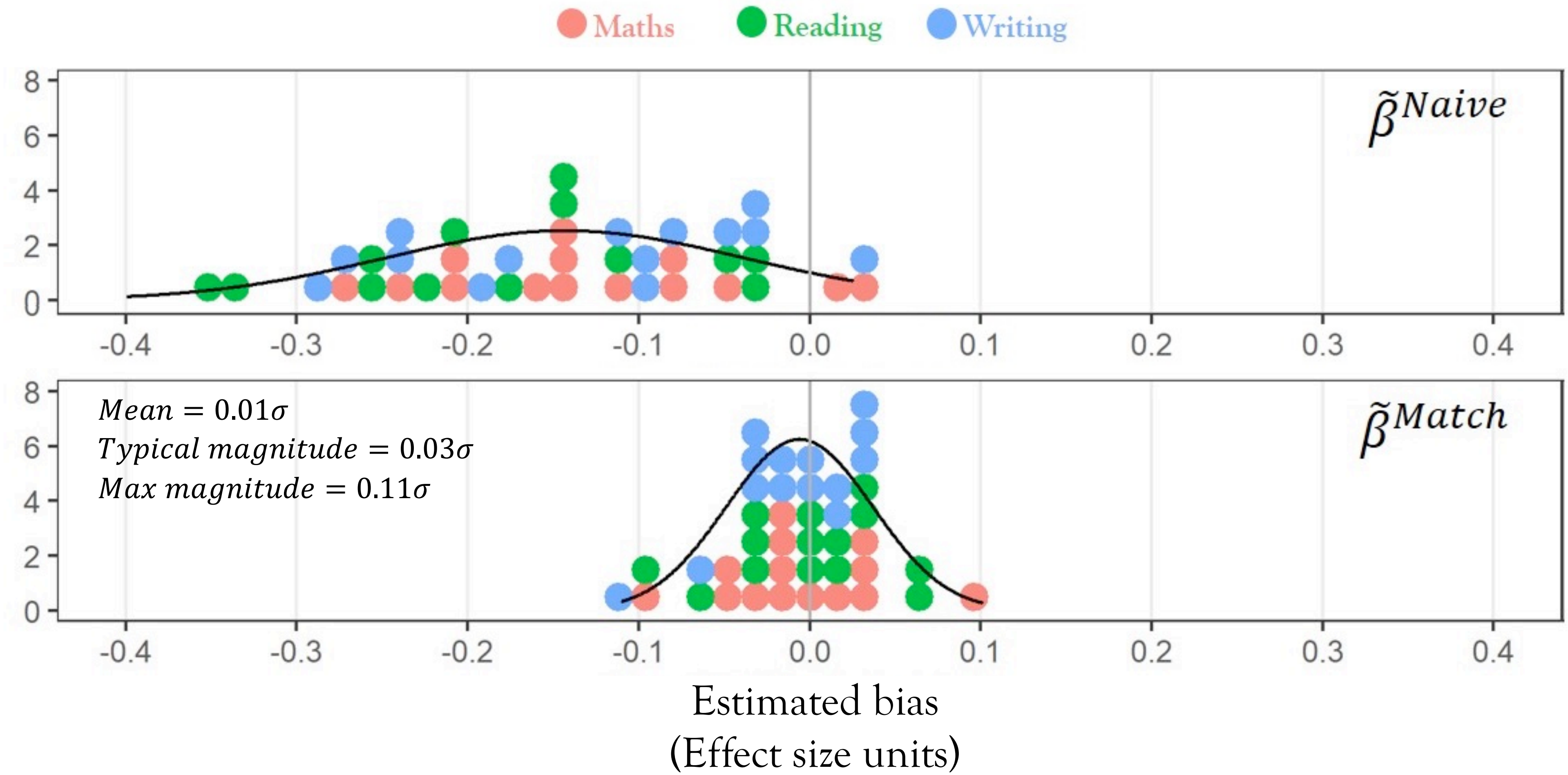
- β^{Naive}
- Simple contrast between RCT control and observational control
- Initial assessment of how big an issue selection bias might be (Wong et al. 2018)

Bias after conditioning on observables

- β^{Match}
- Condition on observables
- For each program, we generate a matched comparison group:
  - 1:1 matching
  - No replacement
  - Mahalanobis distance+propensity score caliper
- Our goal was to use a method that:
  - is common in applied research, rather than something cutting edge
  - is computationally cheap (for simulation-based inference)


ttds
mb
p4c
chs
lrs
hh
dt
cm
tott
am
mtg
ref
fl
sm
$\hat{\beta}^{Naive}$
$\hat{\beta}^{Match}$
Maths
Reading
Writing
-0.4
-0.2
0
0.2
0.4
Selection bias (effect size units)

Estimates of underlying bias



Summary of research

- We compared RCTs to Matching 42 times, and didn't find systematic differences.
  - Results were similar for Maths, Reading and Writing outcomes

- Some may be tempted to conclude that "selection bias" isn't a big problem, so long as we're working on school evaluations, with rich admin data…

- …we argue that this goes too far and that there are limitations to bear in mind:
  - Non-radical interventions
  - Selection bias is a 'moving target' and needs constant re-checking

Thanks

Questions

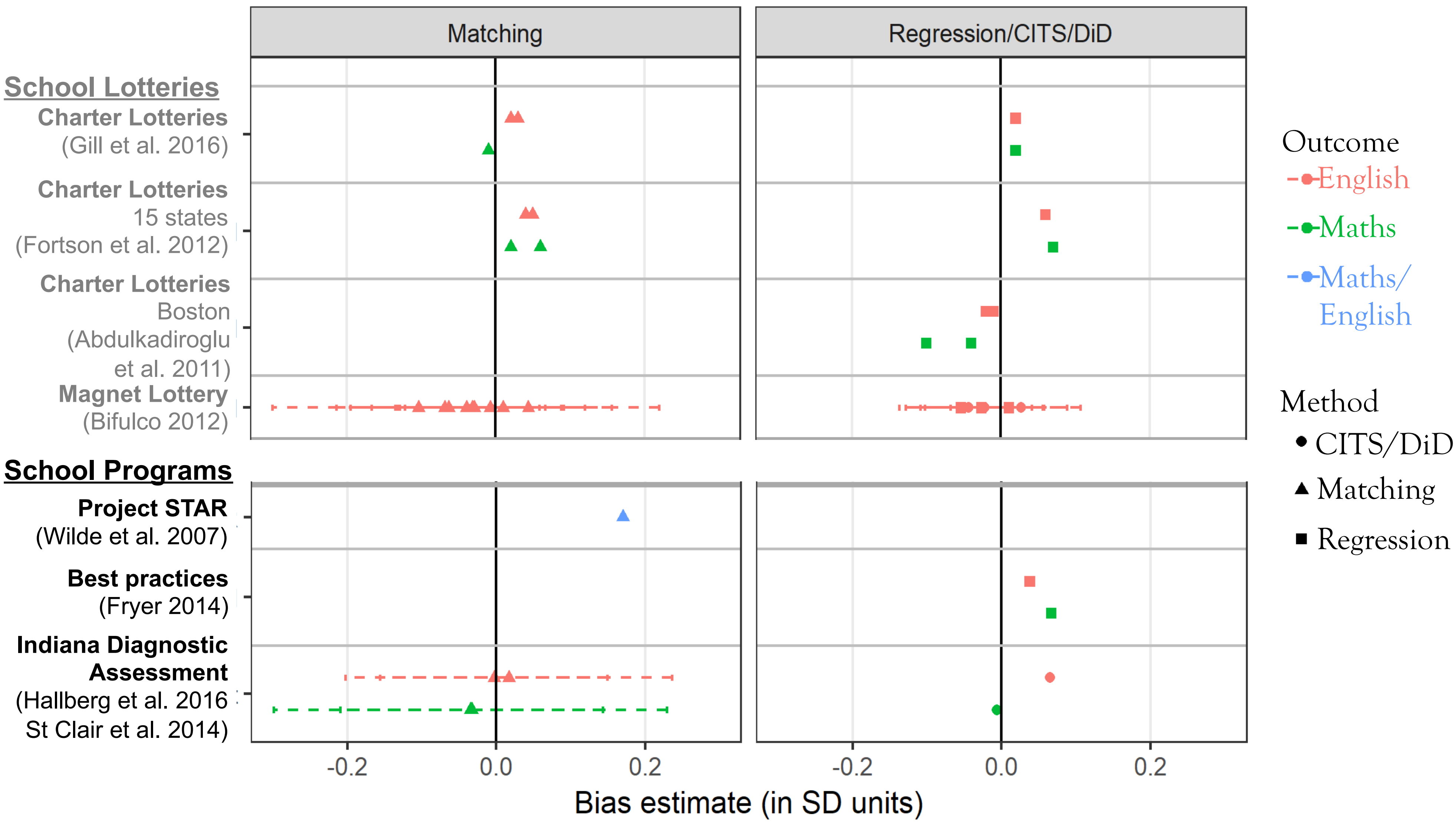
Recommendations for the IES/EEF and researchers (along with some ideas for future work)

1. We should do more observational evaluations using resources like the National Pupil Database in England

2. Conduct within-study comparisons as part of follow-up analyses

3. Use within-study comparisons to systematically examine the performance of different non-experimental methods

Existing evidence from schools: not many estimates, but a promising context



Source: Wong et al. (2018)

Summary of covariates

Category	Label	Level	Description°	Source*
Student achievement	Achievement_grade 2	Student	Average achievement in reading and math in Grade 2	NPD (Key Stage Achievement)
	Late	Student	= 1 if student sits standardized exam a year late	NPD (Key Stage Achievement)
	Early	Student	= 1 if student sits standardized exam in a year earlier than expected	NPD (Key Stage Achievement)
Demographics	Age	Student	Age of student in months	NPD (Pupil Census)
	Free school meals	Student	=1 if student currently gets free school meals	NPD (Pupil Census)
	Gender	Student	= 1 if female	NPD (Pupil Census)
Rurality	Metro	Student	= 1 if student lives in metro area	NPD (Pupil Census)
	Small_metro	Student	= 1 if student lives in small metro area	NPD (Pupil Census)
	Rural	Student	= 1 if student lives in very rural area	NPD (Pupil Census)
	Very rural	Student	= 1 if student lives in very rural area	NPD (Pupil Census)
School-level Achievement	School_academic_mean	School	Predicted achievement in reading and math in Grade 6 (pre-year)	Modelled (based on NPD)
	School _academic_growth	School	Ave. annual change in academic achievement in Grade 6 (4 years prior to RCT)	Modelled (based on NPD)
	School _grade_level_growth	School	Ave. annual change in percent of Grade 6 at grade level (4 years prior to RCT)	Modelled (based on NPD)
School size and type	Voluntary_school	School	= 1 if school is a voluntary school (state-funded, often religious)	NPD (School census)
	Academy_sponsor	School	= 1 if school is a sponsored academy	NPD (School census)
	Academy_converter	School	= 1 if school is a converted academy	NPD (School census)
	Other_type	School	= 1 if school type is not described by the types listed above	NPD (School census)
	Ofsted	School	Integer values of 1 (outstanding) to 4 (inadequate)	Ofsted
	School size	School	Total number of students in school in pre-year	NPD (Finance)
	Type_secondary	School	= 1 if secondary school	NPD (School census)
	Type_middle	School	= 1 if school is a middle school	NPD (School census)
	Type_both	School	= 1 if school has primary and high school	NPD (School census)
Budget	Income	School	Total income in pre-year	NPD (Finance)
	Outside budget	School	Pounds spent on outside programs, services, and ICT	NPD (Finance)
Staffing	TA Percent	School	Proportion of staff who are Teacher Assistants	NPD (Workforce)
	Teacher pupil ratio	School	Pupil teacher ratio in pre-year	NPD (Workforce)
Location variables	Crime	LSOA*	Index of crime	English Indices of Deprivation
	Housing	LSOA*	Index of housing quality	English Indices of Deprivation
	IDACI*	LSOA*	Omnibus index of disadvantage	English Indices of Deprivation

*NPD = National Pupil Database; IDACI = Income Deprivation Affecting Children Index; LSOA = Lower Super Output Area (census region). See Appendix B for details. Pre-year is the year before the RCT randomisation.

Characterising Bias

First, define $E\bar{Y}_{CO}^{adj}$ as the adjusted mean comparison outcome:

$$E\bar{Y}_{CO}^{adj} = \int \mu_{CO}(x) dF_{CT}(x)$$

Where $\mu_{CO}(x) = E[Y(0)|X = x, T = 0, S = 0]$

Now, $\beta = E[\hat{\beta}]$

$$= E[\bar{Y}_{CT}] - \mathrm{E}[\bar{Y}_{CO}]$$

$$= E[\bar{Y}_{CT}] - E\left[\bar{Y}_{CO}^{adj}\right] + E\left[\bar{Y}_{CO}^{adj}\right] - E[\bar{Y}_{CO}]$$

$$= \Delta_U + \Delta_X$$

Estimating $\boldsymbol{\beta}^{\boldsymbol{Match}}$ in more detail

- β^{Match} is a contrast between RCT control, and matched comparison group

- After generating a matched comparison group, we estimate β^{Match} using a regression model

- For each intervention w and outcome k, we fit the following model for pupil $\boldsymbol{i}$ in school $\boldsymbol{j}$:

$$\mathrm{Y}_{\mathrm{ijkw}} = \alpha_j + \boldsymbol{\gamma X_{ij}} + \beta_{kw}^{Match} S_j + \epsilon_{ijkw}$$

$$\alpha_j \sim N(\alpha_0, \sigma_\alpha^2)$$

$$\epsilon_{ijkw} \sim N(0, \sigma^2)$$

Meta-analysis

- Observed estimates of selection bias, $\hat{\beta}_{kw}$ are modelled as follows:

$$\hat{\beta}_{kw}|\beta_{kw} \sim N\left(\beta_{kw}, \sigma_{kw}^2\right)$$

$$\beta_{kw} \sim N(\nu, \tau^2)$$

Where

- β_{kw} is the underlying bias. We model this as a random effect that differs across interventions and outcomes. The mean bias is ν and the variance is τ^2
- Observed estimates of bias deviate from the underlying parameter due to sampling variation, which is captured by σ_{kw}^2

After estimating $\hat{\tau}^2$ and $\hat{\nu}$ we generate empirical Bayes estimates of bias:

$$\beta_{kw}^* = \hat{\lambda}_{kw}\hat{\nu} + \left(1 - \hat{\lambda}_{kw}\right)\hat{\beta}_{kw}$$

Where: $\hat{\lambda}_{kw} = \frac{\hat{\sigma}_{kw}^2}{\hat{\sigma}_{kw}^2 + \hat{\tau}^2}$

Finally, we turn these into *contrained* empirical Bayes $\tilde{\beta}_{kw}$ so that $var\left(\tilde{\beta}_{kw}\right) = \hat{\tau}^2$

Sources: Higgins, Thompson and Spiegelhalter (2009); Weiss et al. (2017)

Meta-analysis details (part 1)

- We estimate τ using the method of moments approach from Higgins et al. (2009):

$$\hat{\tau}^2 = \max\left\{0, \frac{Q-(K-1)}{\sum\hat{\sigma}_{kw}^{-2} - \frac{\sum\hat{\sigma}_{kw}^{-4}}{\sum\hat{\sigma}_{kw}^{-2}}}\right\}$$

Where:

$$Q = \sum\left(\hat{\beta}_{kw} - \bar{\beta}\right)^2 \hat{\sigma}_{kw}^{-2}$$

$$\bar{\beta} = \frac{\sum\hat{\beta}_{kw} \cdot \hat{\sigma}_{kw}^{-2}}{\sum\hat{\sigma}_{kw}^{-2}}$$

- Then, letting $\widehat{\omega}_{kw} = \left(\hat{\sigma}_{kw}^2 + \hat{\tau}^2\right)^{-1}$, we estimate $\hat{\nu} = \frac{\sum\widehat{\beta}_{kw}\widehat{\omega}_{kw}}{\sum\widehat{\omega}_{kw}}$
- Estimates of $\hat{\sigma}_{kw}^2$ come from our simulations under the null

Meta-analysis details (part 2)

- K is the effective sample size, and is based on the icc of the bias estimates [i.e. the intra-class correlation within cluster, defined as $\hat{\rho}$]

- Specifically:

$$K = \frac{kw}{1-(k-1)\cdot\hat{\rho}} = \frac{42}{1-(3-1)\cdot 0.56} = 19.9$$

Where our estimate of $\hat{\rho}$ comes from a multilevel model in which $\hat{\beta}_{kw} \sim N(\alpha_w, \sigma_e^2)$, $\alpha_w \sim N(\gamma_0, \sigma_a^2)$, and $\hat{\rho} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2}$.

- Once we have estimates of $\hat{\tau}$ and $\hat{\nu}$ we get simple empirical Bayes estimates (shrinkage estimates):

$$\beta_{kw}^* = \hat{\lambda}_{kw}\hat{\nu} + \left(1-\hat{\lambda}_{kw}\right)\hat{\beta}_{kw}$$

$$\hat{\lambda}_{kw} = \frac{\hat{\sigma}_{kw}^2}{\hat{\sigma}_{kw}^2 + \hat{\tau}^2}$$

- We then scale the β_{kw}^* so that $var(\beta_{kw}^*) = \hat{\tau}^2$

Meta-analysis: sensitivity check

- There don't appear to be differences across outcomes (maths, reading writing)
- So, as a sensitivity check, we re-run our meta-analysis treating each intervention as 1-unit

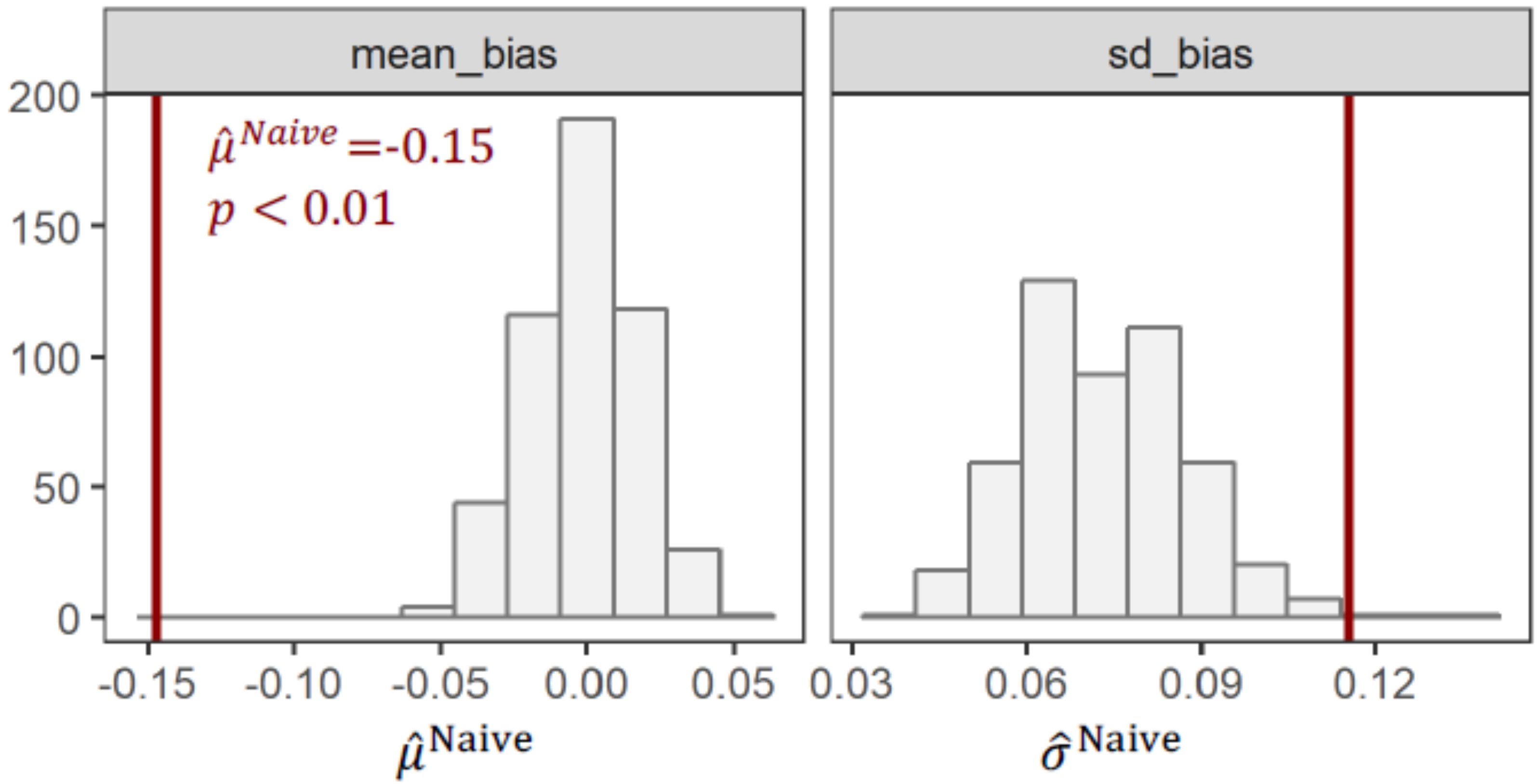
$$\hat{\beta}_w = \frac{1}{3}\Sigma\hat{\beta}_{kw} \qquad and \qquad \hat{\sigma}_w = \frac{1}{3}\Sigma\hat{\sigma}_{kw}$$

- We use the same Method-of-Moments approach (but now K=14).
- The estimated value of Q is smaller than $(K-1)$, so the MoM estimate of $\hat{\tau}^2$ at the intervention level defaults zero, as

$$\hat{\tau}^2 = \max\left\{0, \frac{Q-(K-1)}{\Sigma\hat{\sigma}_w^{-2} - \frac{\Sigma\hat{\sigma}_w^{-4}}{\Sigma\hat{\sigma}_w^{-2}}}\right\}$$

- The 95 percent confidence interval of $\hat{\tau}^2$ using is [0,0.05], which we generate using test inversion (a la Weiss 2017)

Null hypothesis testing



Null hypothesis testing