Bigger Data and stronger Causal Inference from Quasi-Experiments

Thomas D Cook
Northwestern University and Mathematica, Inc
University of Connecticut, 2015

Any one Bigger Dataset entails…

- More variables, hence more constructs
- More versions of same construct --more reliability
- More frequent assessments – more time series
- Denser sampling – more cases
- More local sampling – better comparisons
- ALL are features of increased dimensionality – more constructs, more times and periods, more and more local comparison opportunities

The Great Amplifier: Linkages

- With identifier – SS #, tel., fidelity card, face recognition software – can link data across data sets
- For same individuals or aggregates like households and schools and n'hoods
- Identifiers can themselves be linked
- They multiply dimensionality

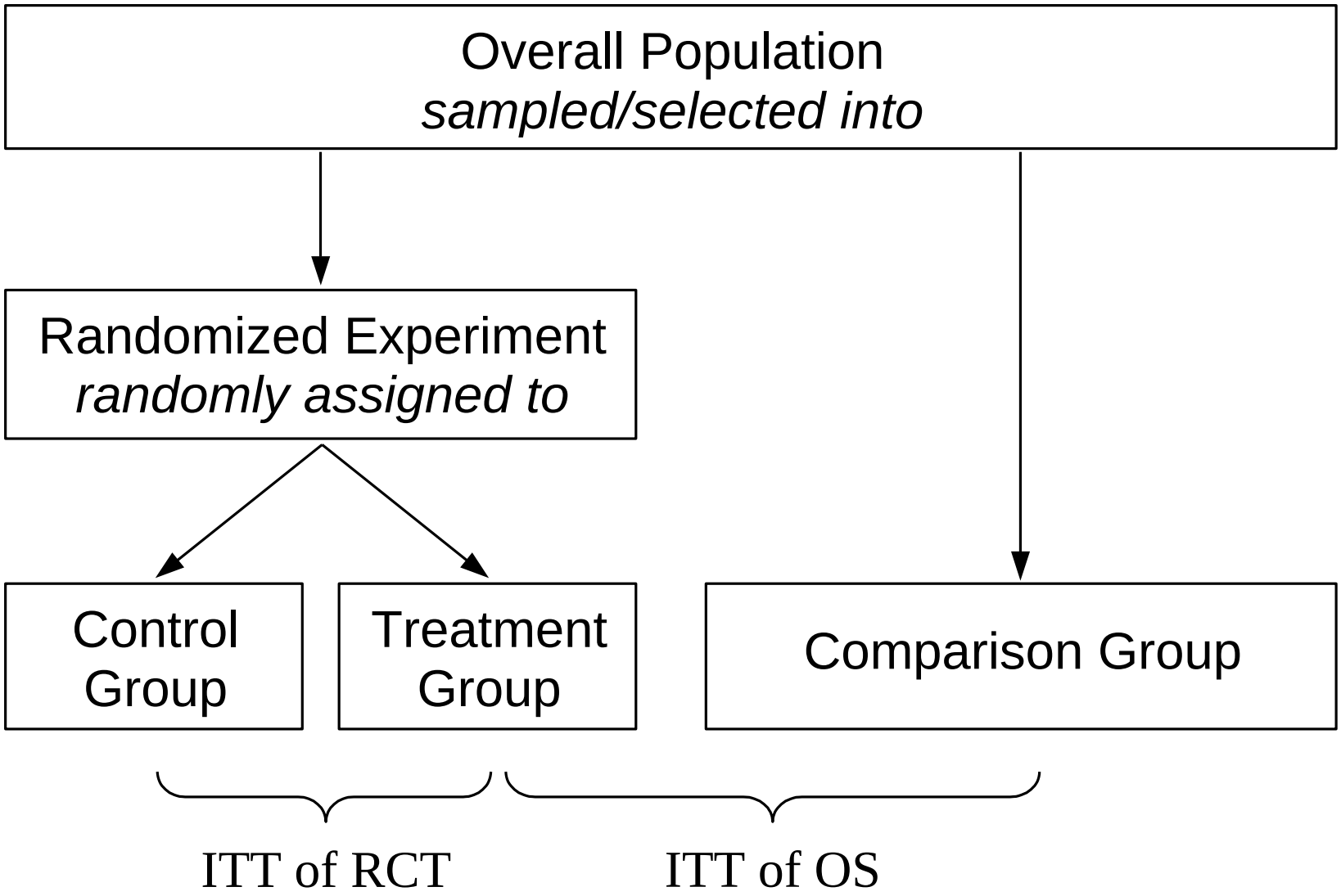
Fast Forward to Today

- Google, Facebook etc. as Big Datasets
- NSA as ever more linked data
- Data on multiple kinds of constructs
- Some “nudge-like” RCTs to improve practice
- Use other methodologies too, esp time series
- Perhaps someone even examines whether the nudge RCTs and ITS give similar results, or conditions under which they do

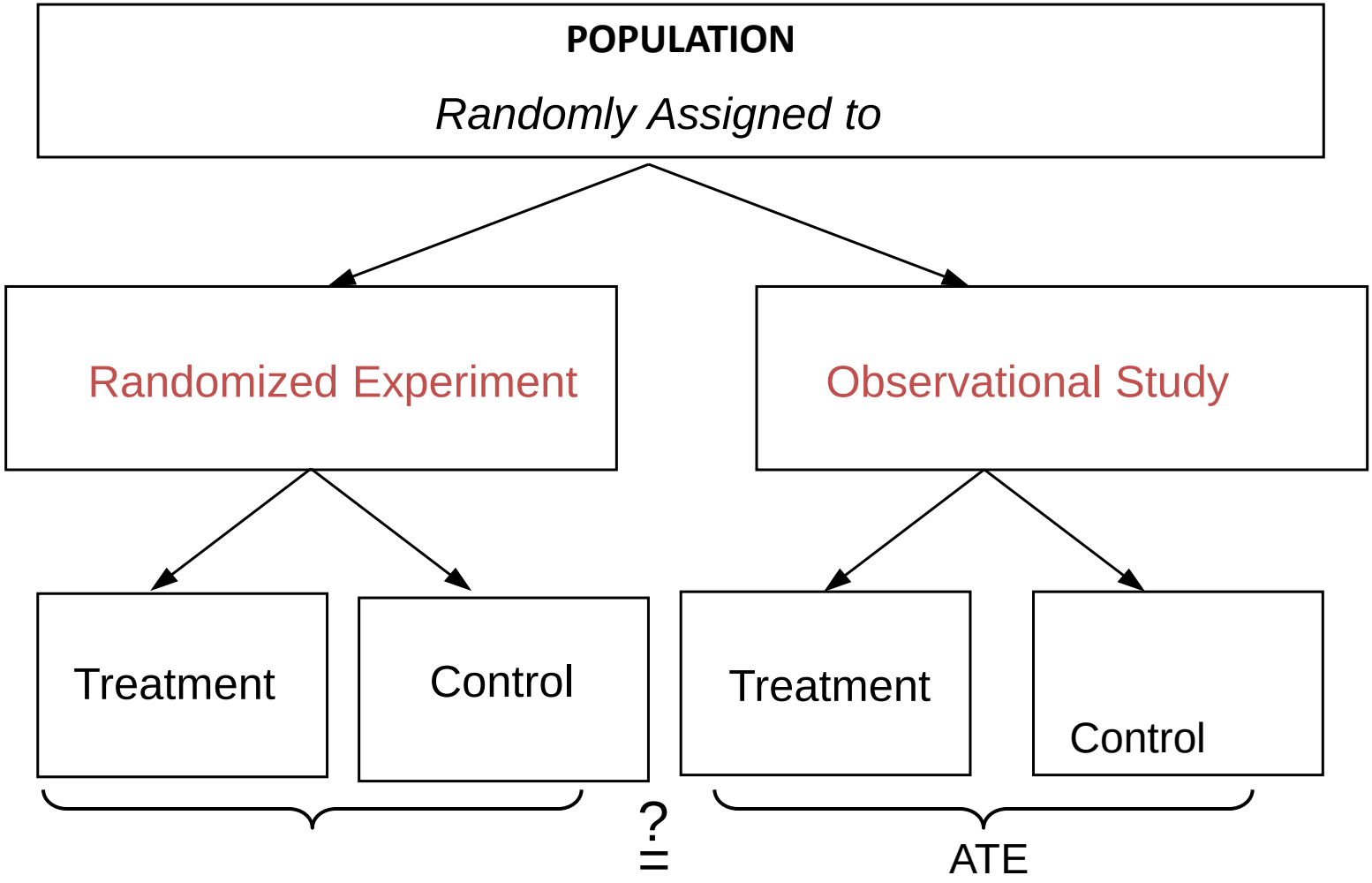
Our concern today with

- Evaluation of exogenous shocks, like programs but not like low grade nudges
- Assuming bigger data does little to improve probability of RA, how might evaluation be improved by “better” comparisons, by more time points, by a single pretest time point, by higher reliability, by just more data?
- But how do you know the causal answers are improved, and to an acceptable degree?

WSC Design: Three-Arm Study



WSC Design: Four-Arm Study



Within-Study Comparison aka Design Experiment

- Have a benchmark– RCT in 62 of the 70 examples to date. Compute causal estimate and SE
- Quasi-experiment of many different “types” – CITS, RD, NECGD with pretest and local match, NECGD with fully known selection process, with partially known, with scarcely known etc. –
- Attach same treatment group to RCT and QE, adjust QE, and then compute QE estimate
- Compare QE and RCT estimates, and conclude

Conditions for a good WSC

- A well implemented RCT, with minimal sampling error
- No third variable confounds – like from measurement
- Comparable estimands – RD and RCT
- Blinding to the RCT or adjusted QE results
- Defensible criterion for correspondence of RCT and adjusted QE results

Limitations of WSCs

- Only be done on topics with benchmark
- No reason to believe that a given QE will always replicate RCT finding; goal is to identify designs that often replicate findings.
- This is inductive and requires a large sample of WSCs. This talk is not the final word. Even more WSCs needed.

SELECTING A NON-EQUIVALENT COMPARISON GROUP IN QE TO REDUCE INITIAL NON-EQUIVALENCE

The Trick with most QEs is

- To select an intact C group as similar to T as possible to minimize selection difference thru sampling. Contrast is with making them seem similar through individual case matching
- To use covariates in analysis that reduce any selection difference still remaining. This is where propensity scores, ANCOVA come in.
- Heckman advice: Local comparison groups plus pretest measure of the outcome

What does Local “Mean”?

- Identical twins, non-identical, sibs, cousins
- Same grade cohort in schools, birth cohort
- Schools in same district vs other
- Job training sites in same local labor market
- Towns at border of different states vs all state
- More local the better since it matches on more unobservables as well as observables

Local intact comparison groups

- Past empirical research in Cook et al. (2008) shows 3 cases in different fields where local choice eliminated all bias. Two more WSCs since, and two others earlier with same result.
- But some counter-cases in job training. Always reduces bias but does not always eliminate it
- Problem is: Not all local matches are good
- How to shift the odds if data sufficiently "rich"?

Focal Matching

- Seeks to make selection strongly ignorable
- Use covariate measures that tap into all those factors that are (a) correlated with selection into T and correlated with outcome/effect
- You rarely know these, and so best guesstimate thereof, incl direct study
- This kind of matching a spotty record of recreating RCT impact except where considerable effort goes into learning selection

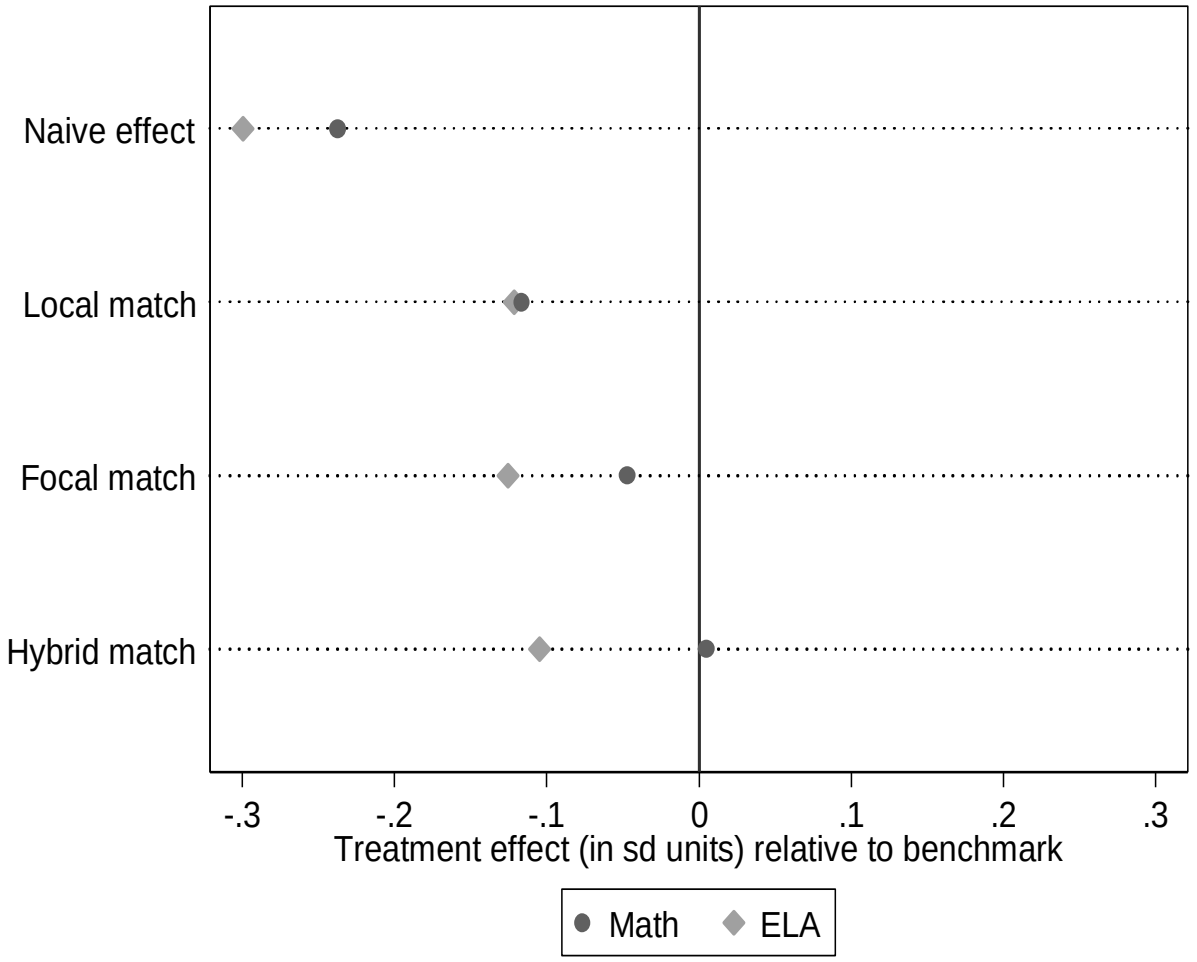
Hybrid sampling model of Stuart and Rubin (2008)

- Define caliper for adequacy of a match
- Match all LOCAL Cs to T that fall within caliper
- For others, perform a match using a PS predicated on analysis of selection processes
- Result = mix of acceptably matched local Cs that control for more unobservables, and acceptably matched non-local Cs, but matched only on observables

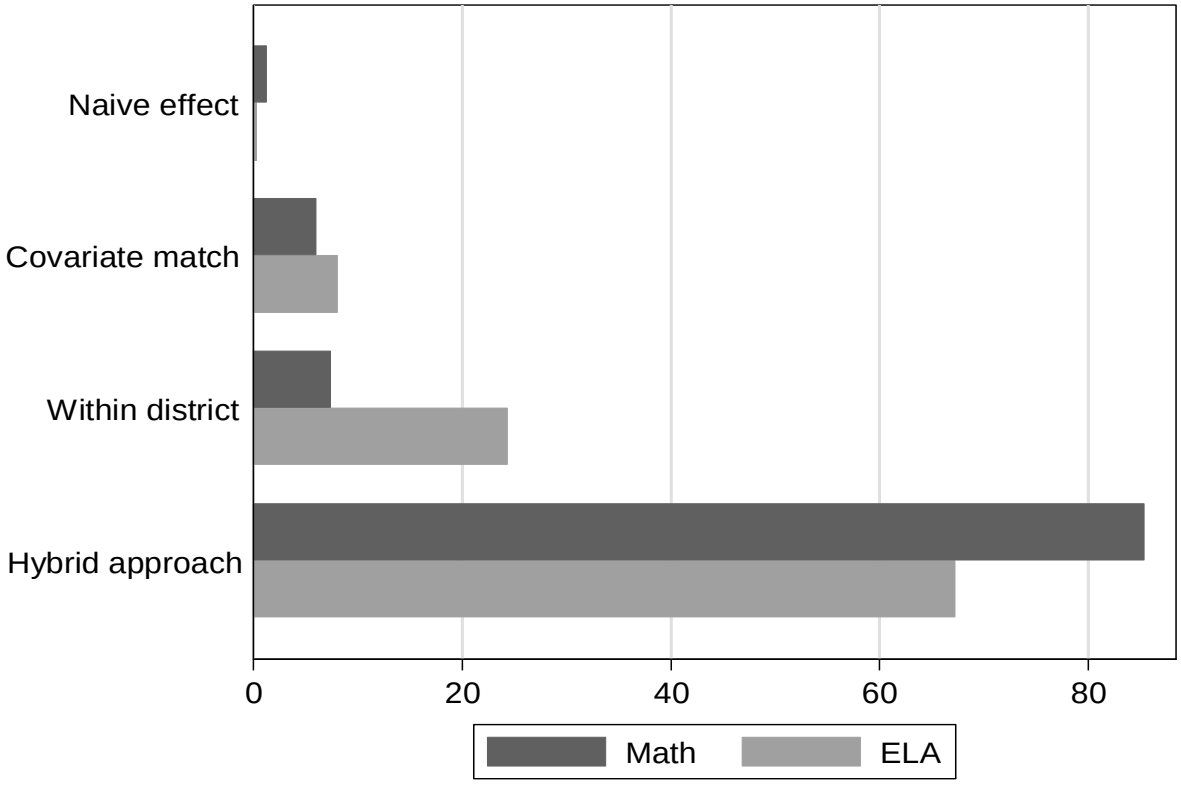
Hallberg, Wong, & Cook (in press)

- This paper draws on a WSC to examines correspondence with the RCT benchmark (Indiana student feedback study) after matching
  - Within district as long as the schools do not differ by more than 0.75 standard deviations of the propensity score (Local)
  - For others match on observed school-level covariates known to be highly correlated with the outcome of interest (Focal)
  - Combine both T and C matched cases (Hybrid)

Performance of local, focal and hybrid matching across two dependent variables



Percentage of times observational approach performed best across 1000 replications



Summary

- Need for more studies of hybrid matching
- Intact group matching increases overlap. Useful first stage in a QE design strategy?
- Local matching matching is always useful and often brings about RCT result.
- Neither is a guarantee, like well implemented RCT would be

MORE PRETEST DATA POINTS: RCT VS. INTERRUPTED TIME SERIES (ITS) AND ESPECIALLY COMPARATIVE INTERRUPTED TIME SERIES (CITS)

Interrupted Time Series Can Provide Strong Evidence for Causal Effects

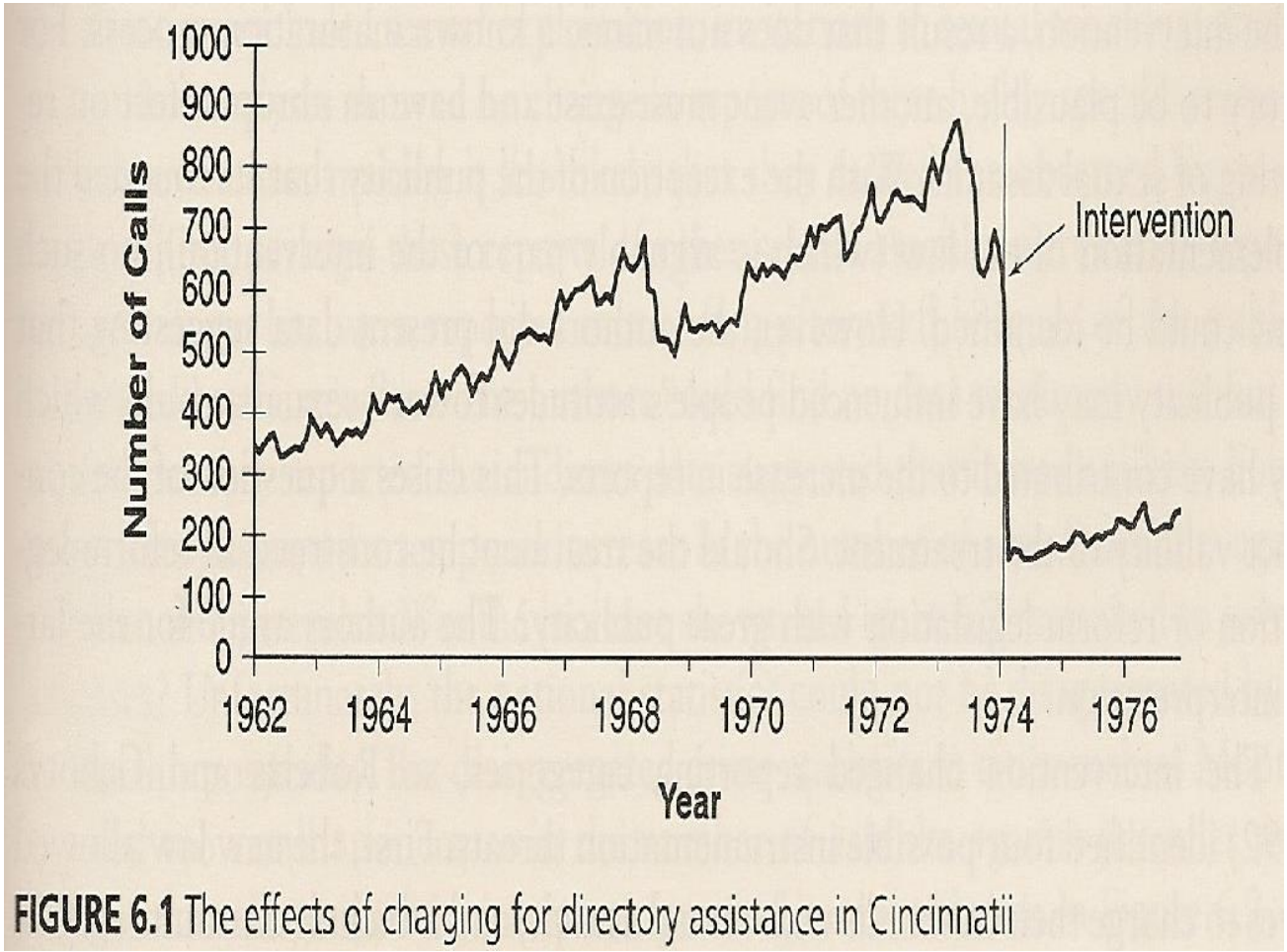


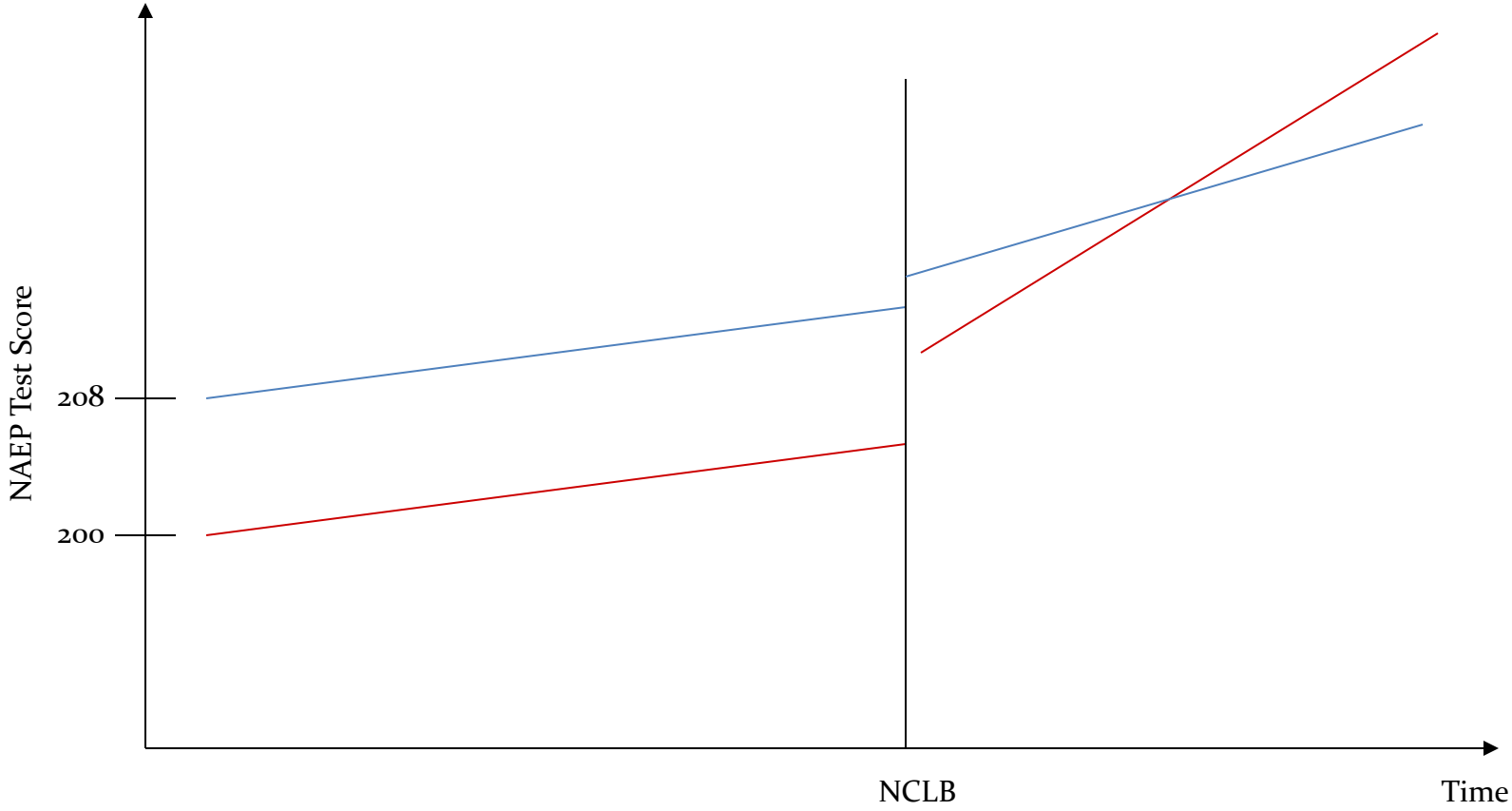
FIGURE 6.1 The effects of charging for directory assistance in Cincinnatii

- Clear Intervention Time Point
- Huge and Immediate Effect
- Clear Pretest Functional Form + many Observations
- No Alternatlve at Interventio Can Explain Change

Limitations of Simple One-Group ITS

- History, around the intervention point
- Instrumentation
- Stat Regression
- Functional form extrapolation needed
- Analysis has to account for correlated errors (we will not deal with this issue here)
- Suggest the advisability of a comparative ITS

Hypothetical NCLB effects on public (red) versus private schools (blue)



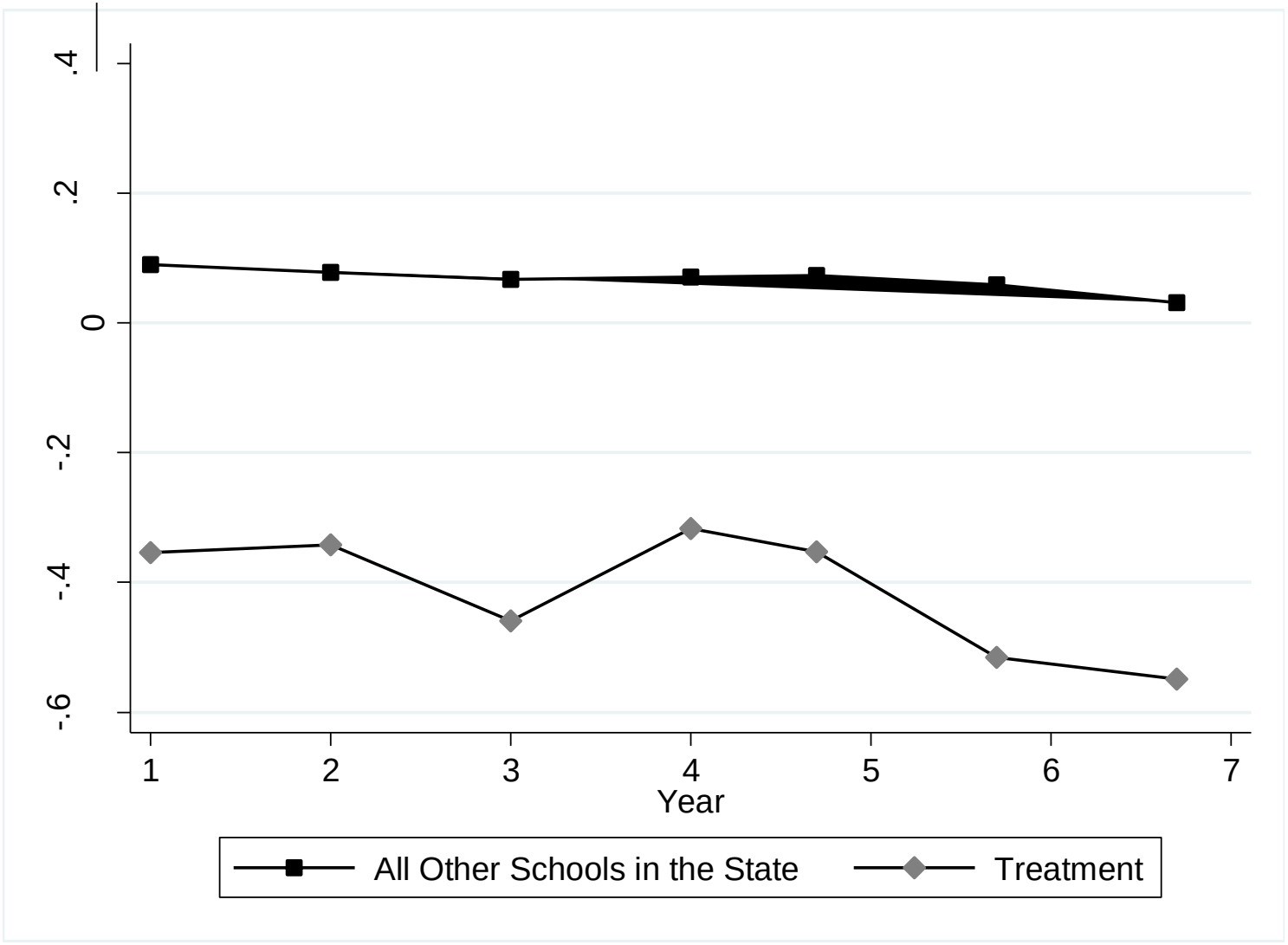
WSC and CITS

- Six studies in medicine, four in education, one in environmental sciences
- All claim causal inferences similar
- No meta-analysis to date
- No analysis of file drawer problem
- Remarkable cos these internal validity threats could have operated but did not

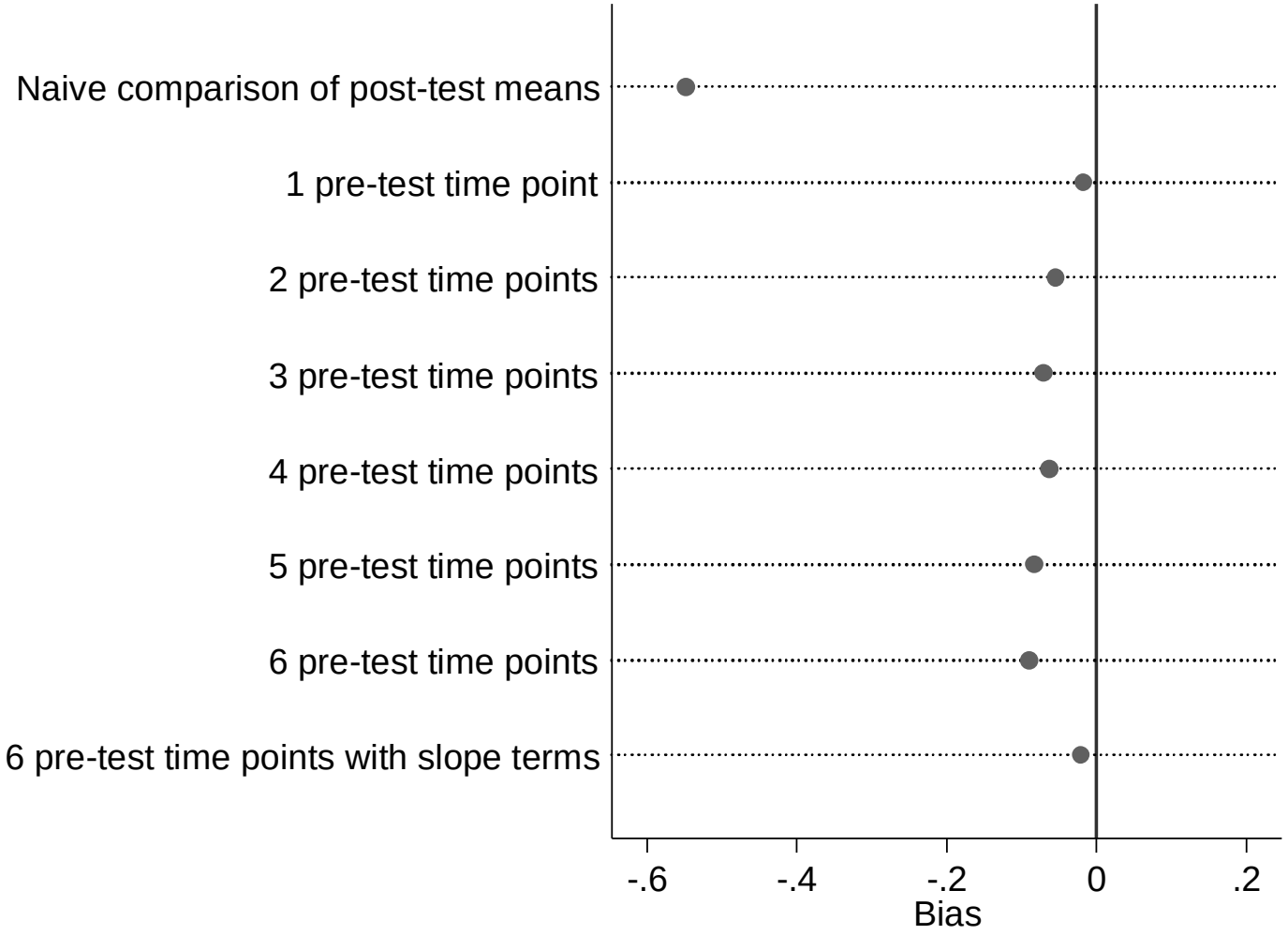
St. Clair, Cook, & Hallberg (2014)

- RCT: Study of Indiana's system for feedback on student performance (schools as unit of assignment)
- Comparative ITS comparison groups
  - Basically all schools in the state
  - Matched schools in the state

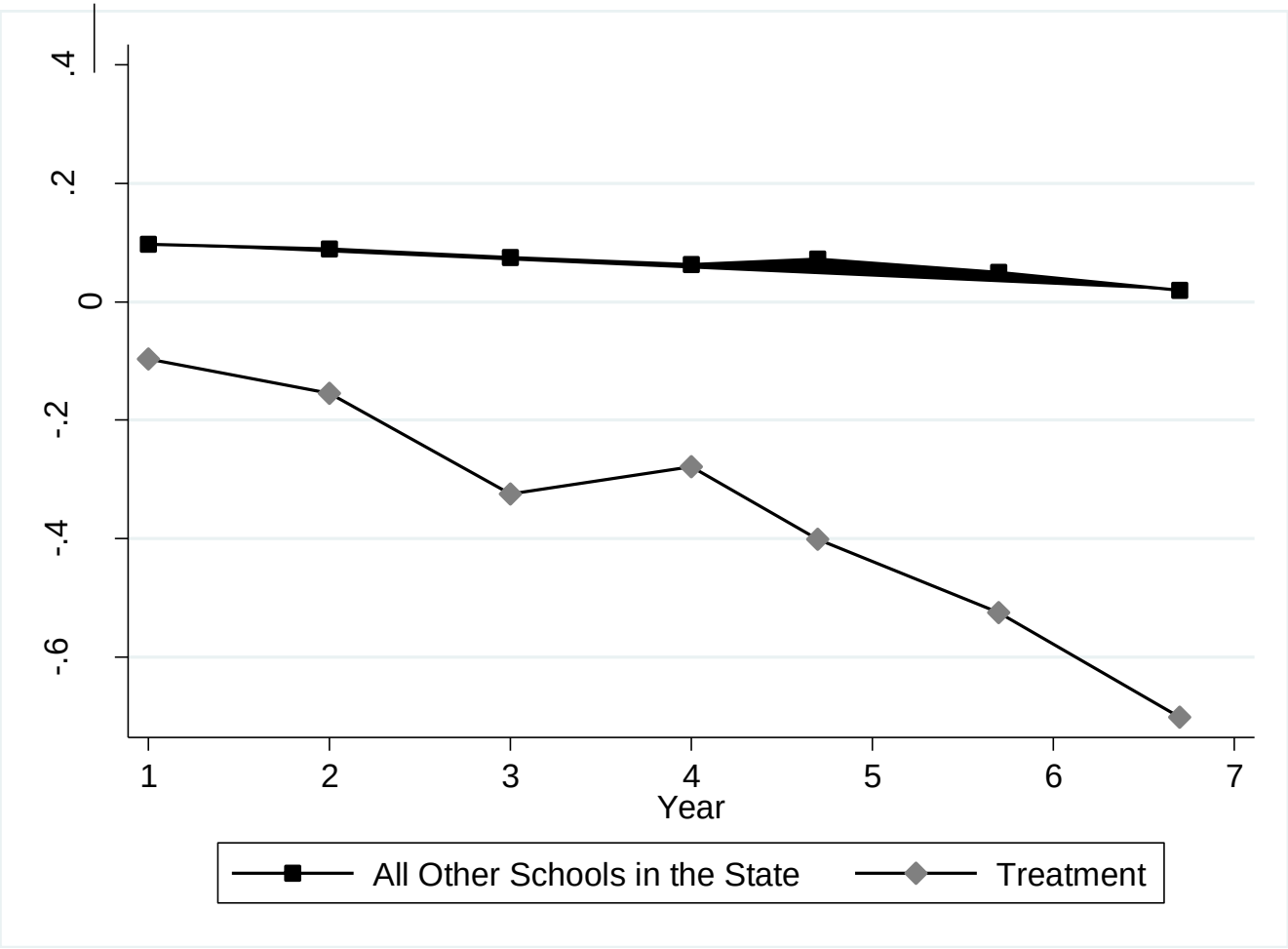
Math (All schools)



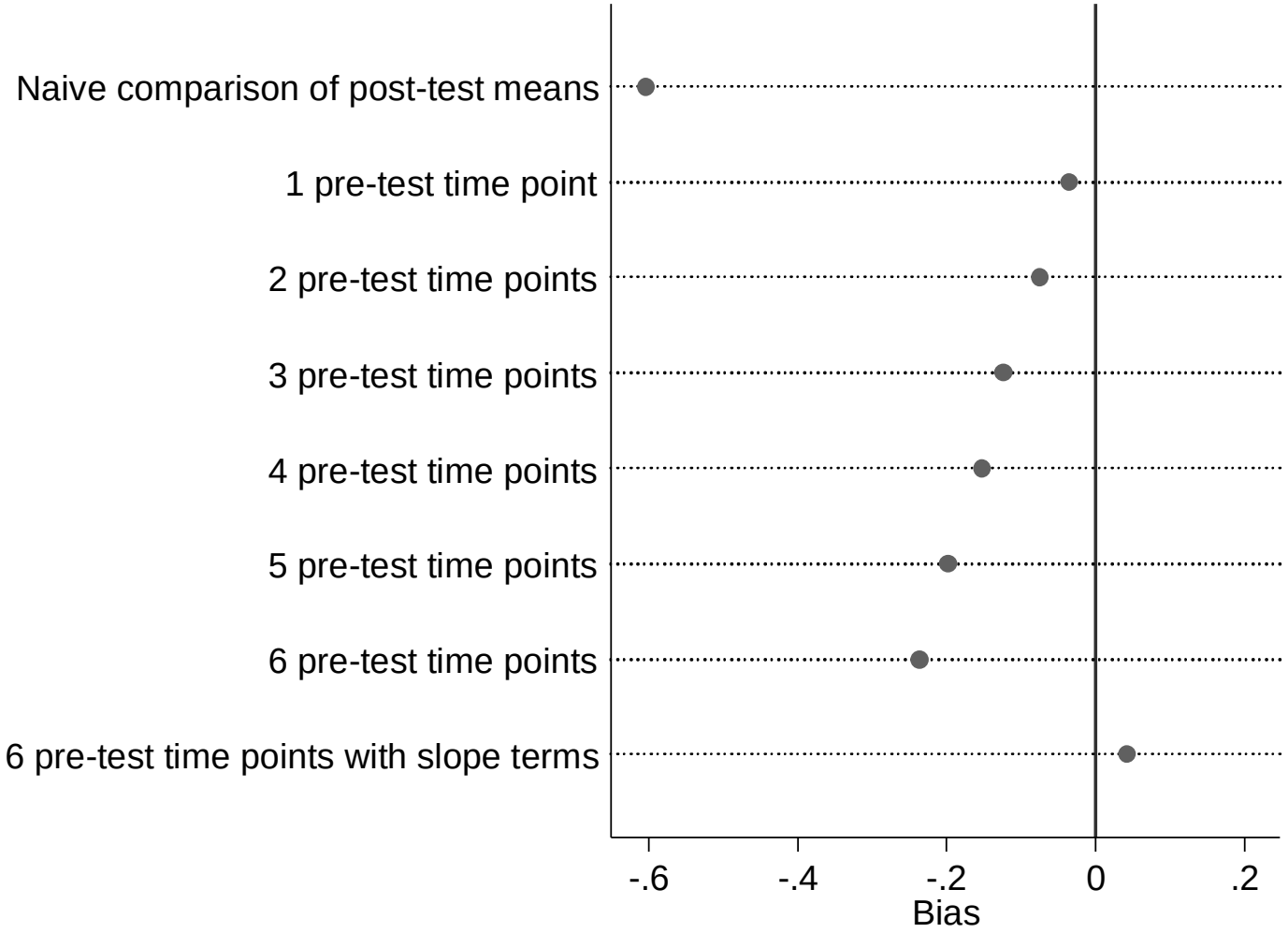
Math: WSC Results



ELA (All Schools)



ELA: WSC Results



What about Matching C to T Units?

- We can match C to T units, though this entails some case loss. Then no need to assume functional form is correct
- Same results
- Somers et al got the same results
- Environmental science found replicate RCT only with matching
- Matching safest analysis unless sure of FF

CITS Summary

- To date, CITS does well relative to RCT Matching is the most consistent to date
- Models with the correct functional form do well; and one can observe the functional form
- Similar effects despite possible group differences in (a) pre-treatment trend,(b) historical events at treatment; (c) changes in instrument; (d) stat regression– have never been confounds

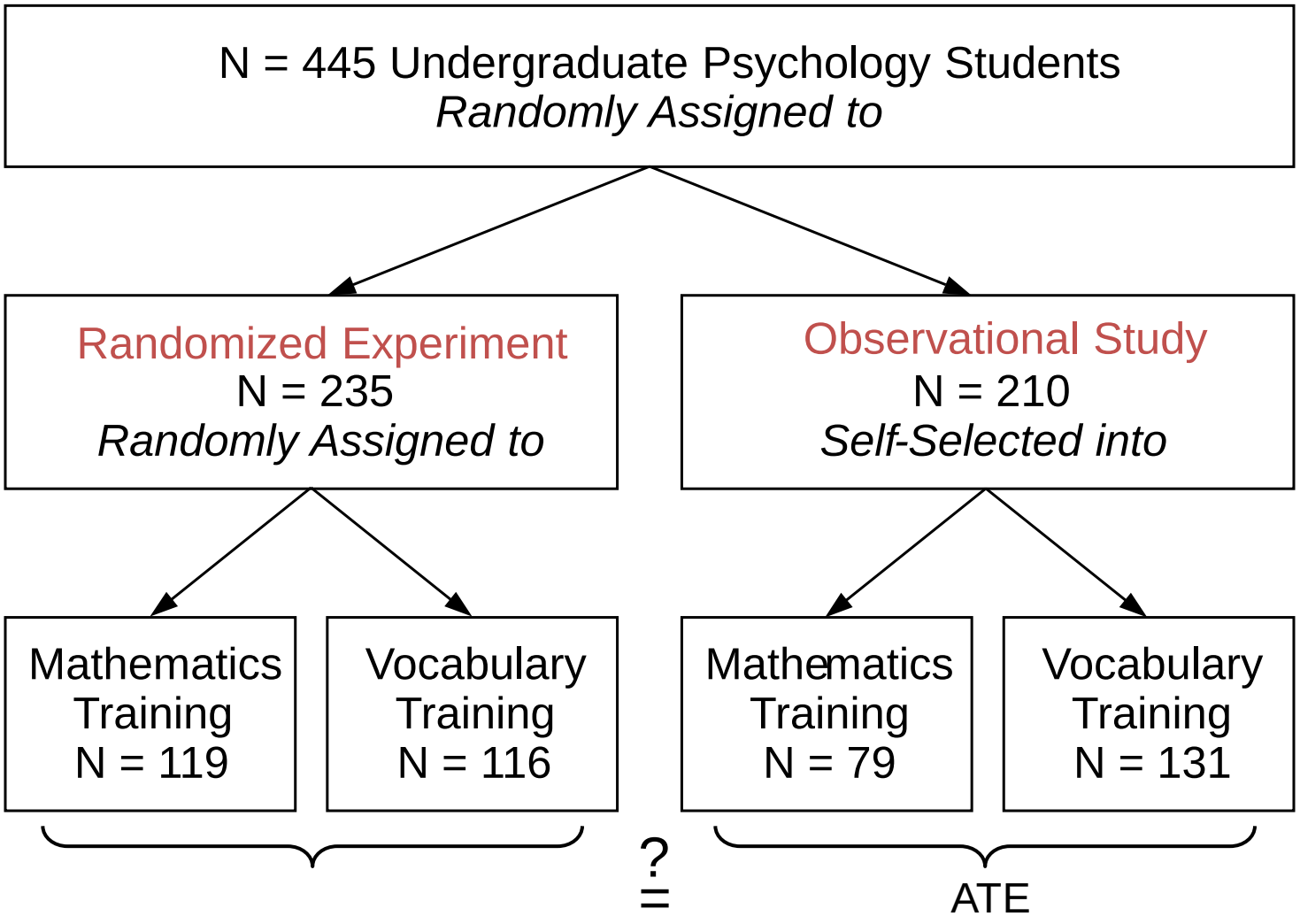
MORE COVARIATES FOR MODELING A STRONGLY SUSPECTED SELECTION PROCESS

Statistical Theory

- Knowing selection and measuring it perfectly gives unbiased causal inference
- BUT rarely know it fully – RDD exception
- Yet we often know major elements of selection – why children are retained in grade; why couples self-select into divorce;
- Here's one example – why students self-select into learning English or math

Strongly suspected selection process

Shadish, Clark & Steiner (2008)



23 Constructs and 5 Construct Domains assessed prior to Intervention

Proxy-pretests (2 multi-item constructs):

36-item Vocabulary Test II, 15-item Arithmetic Aptitude Test

- *Prior academic achievement* (3 multi-item constructs):
  High school GPA, current college GPA, ACT college admission score
- *Topic preference* (6 multi-item constructs):
  Liking literature, liking mathematics, preferring mathematics over literature, number of prior mathematics courses, major field of study (math-intensive or not), 25-item mathematics anxiety scale

Construct Domains

- *Psychological predisposition* (6 multi-item constructs):

  Big five personality factors (50 items on extroversion, emotional stability, agreeableness, openness to experience, conscientiousness), Short Beck Depression Inventory (13 items)

- *Demographics* (5 single-item constructs):

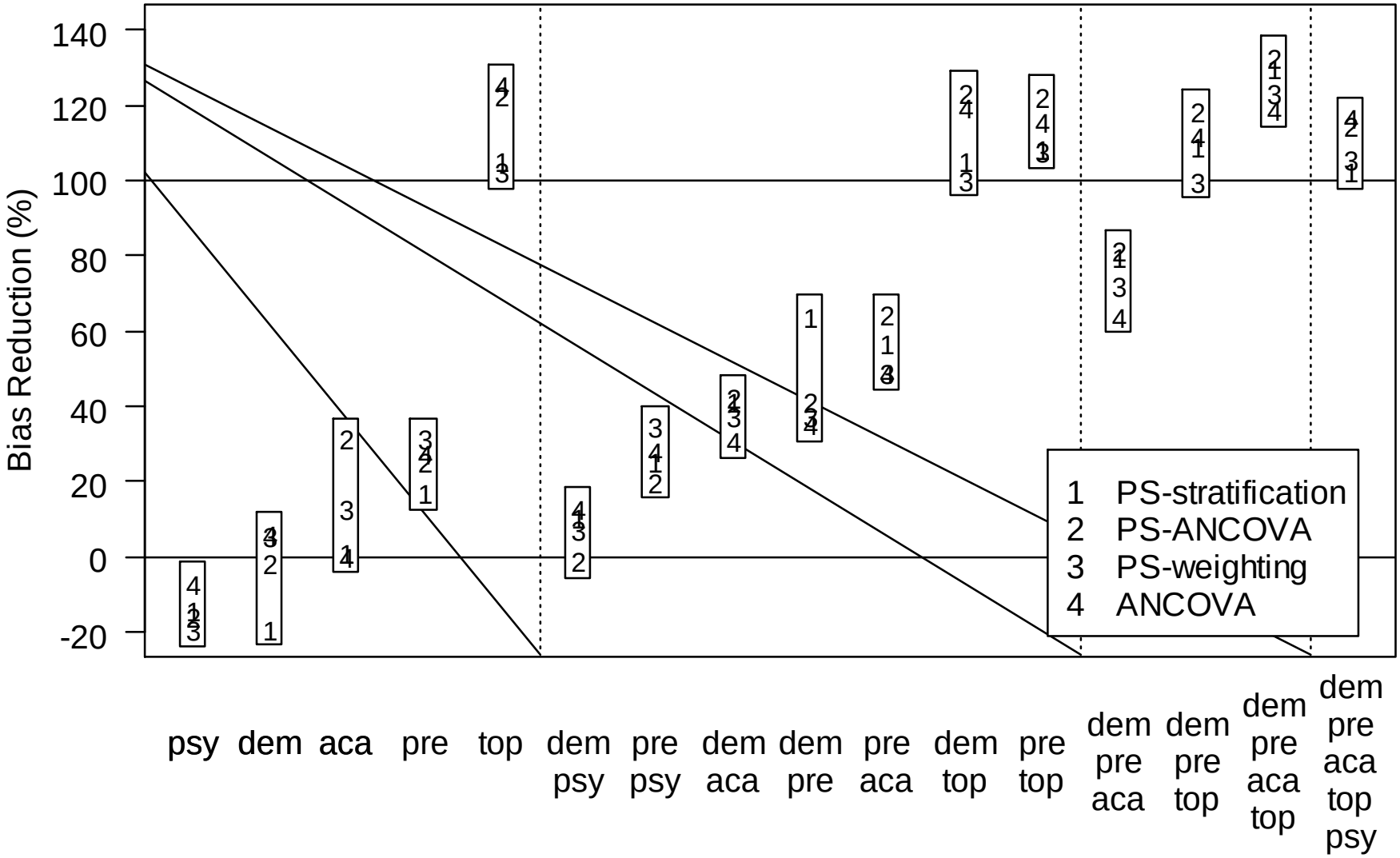
  Student‘s age, sex, race (Caucasian, Afro-American, Hispanic), marital status, credit hours

Was there Bias in the QE with Self-Selection into Tracks?

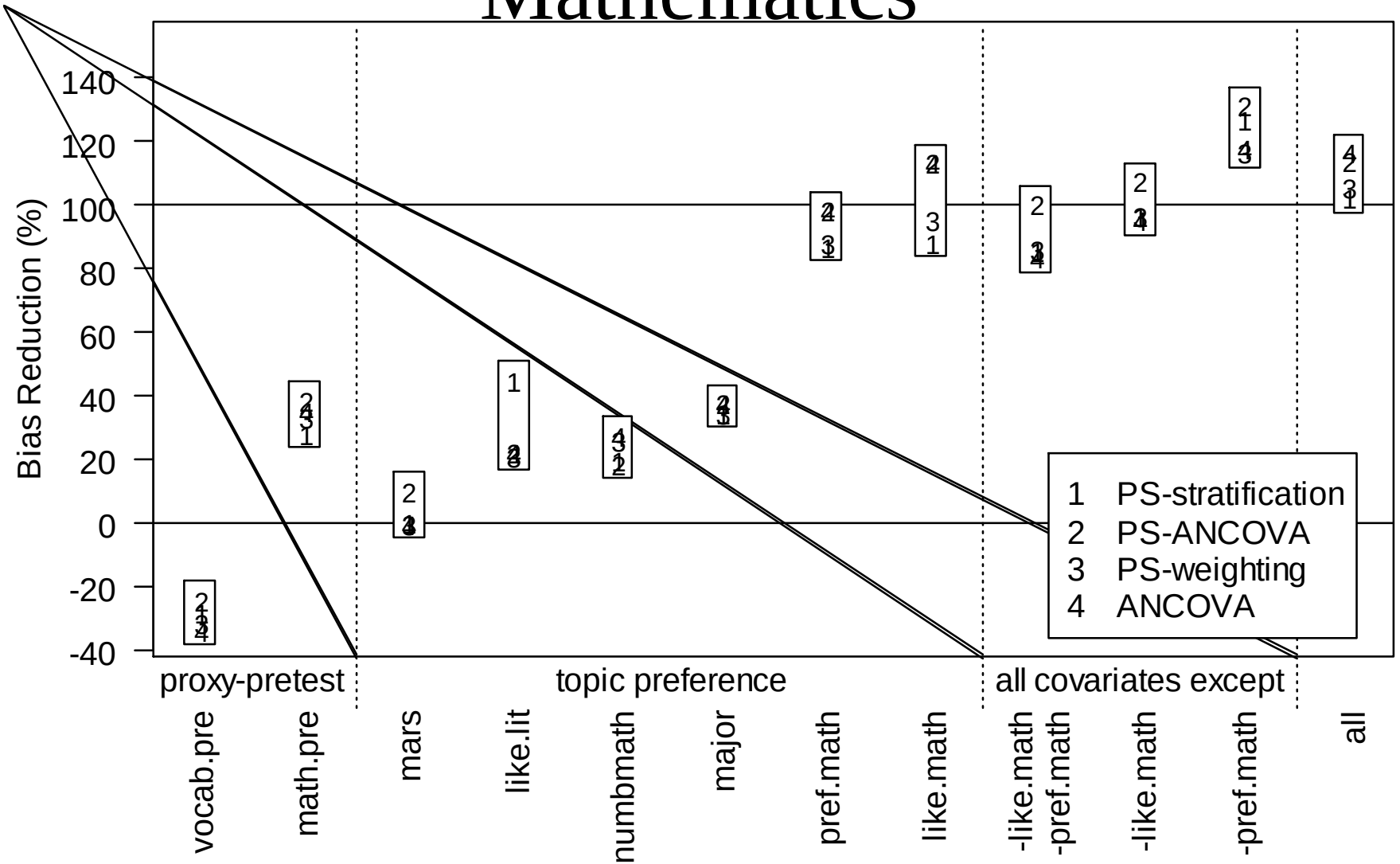
- Random assignment showed effects for each outcome.
- But both math and vocab effects were larger when students self-selected into T versus C
- So our question is: How much of this self-selection bias is reduced by use of covariates measuring several different possible selection processes?

Bias Reduction: **Construct Domains**

Mathematics

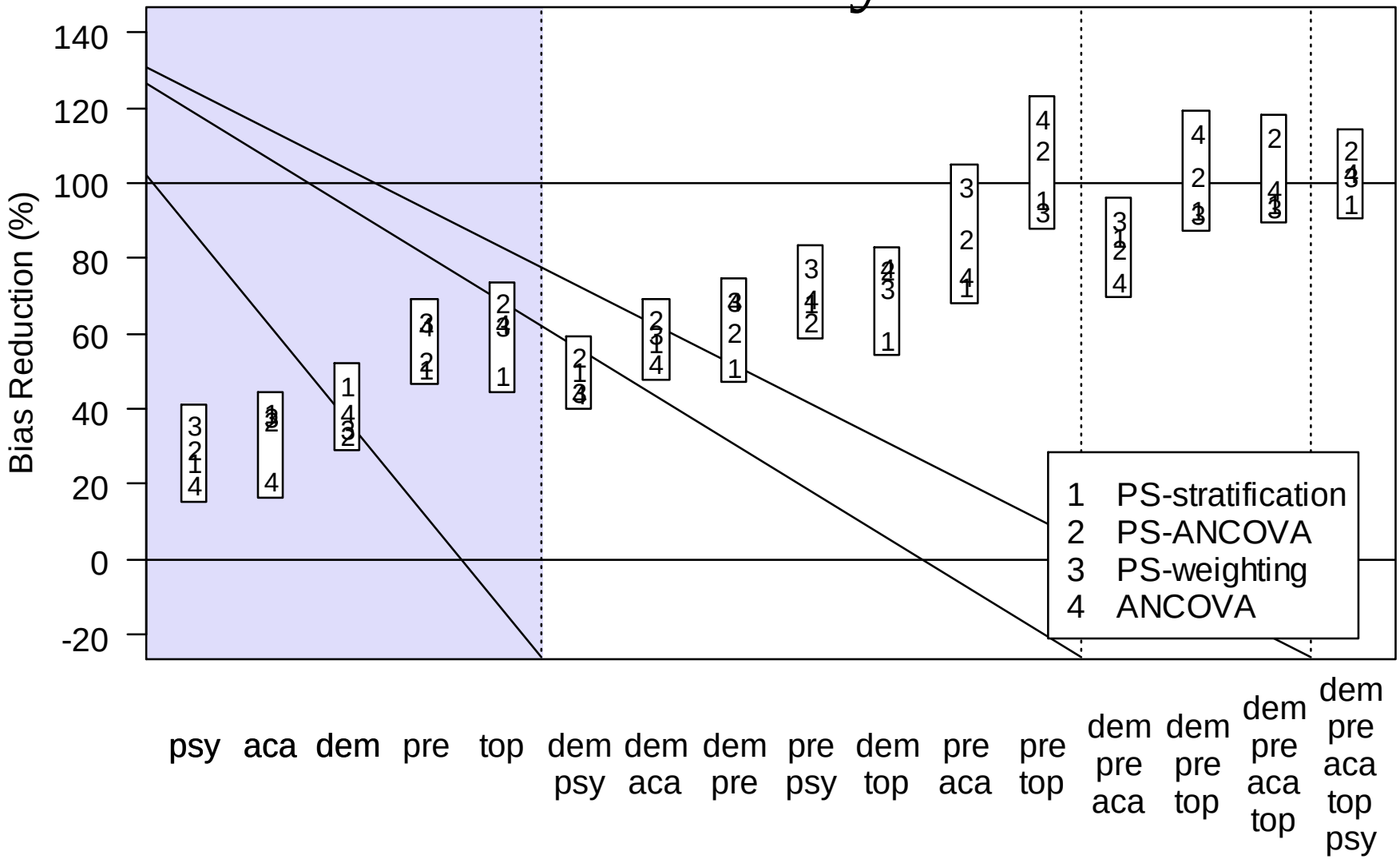


Bias Reduction: **Single Constructs** Mathematics



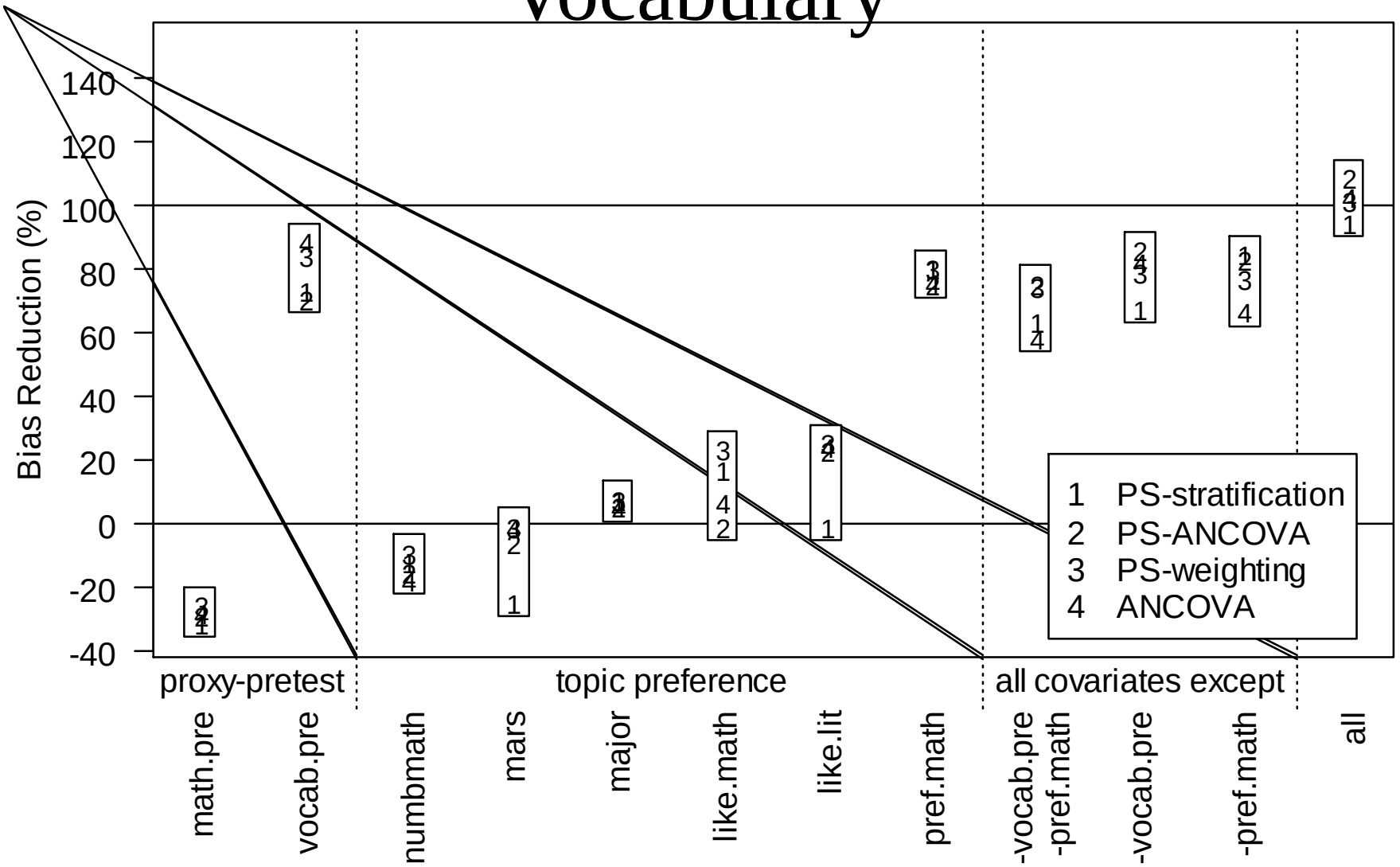
Bias Reduction: **Construct Domains**

Vocabulary



Bias Reduction: **Single Constructs**

Vocabulary



Given Initial Group Differences

- 1. Choice of covariates is crucial
- 2. Reliability counts, but secondary within bounds of 1 to .60.
- 3. Mode of analyzing covariates (OLS and PS matching) makes little difference, though PS preferred in theory
- 4. Replicated in Pohl et al. (2011)

AMONG COVARIATES, HOW SPECIAL IS A PRETEST MEASURE OF STUDY OUTCOME FOR BIAS REDUCTION?

Claims about Pretest

- Claim that pretest is privileged for bias reduction; yet by itself did little for math in Shadish et al.
- In studies modeling the outcome only, pretest often the most highly correlated single variable
- but issue is cor of pretest with selection into T
- Though we suspect selection on pretest to be very frequent, not know how often and when
- Next WSC studies vary when the pretest does and does not vary with selection

Existing Empirical Evidence

- WSCs support privileging true pretest because it is better than others at reducing bias,
- Sometimes reduces all by itself -- Magnet school study (Bifulco, 2010) and earlier CITS studies here
- But it does not always reduce all bias – e.g., Shadish et al. and workforce development lit
- This study examines bias reduction due to pretest *when we vary the correlation with selection* ***both between and within studies***

Between-Studies: Kindergarten Retention

- Hong and Raudenbush (2005; 2006) used rich covariates in ECLS-K to estimate the effect of kindergarten retention on math and reading
- Two prior waves
- Evidence of selection-maturation: Retained have lower mean ***and*** lower rate of change.
- Selection process largely known: past perf and teacher ratigs of performance –both available

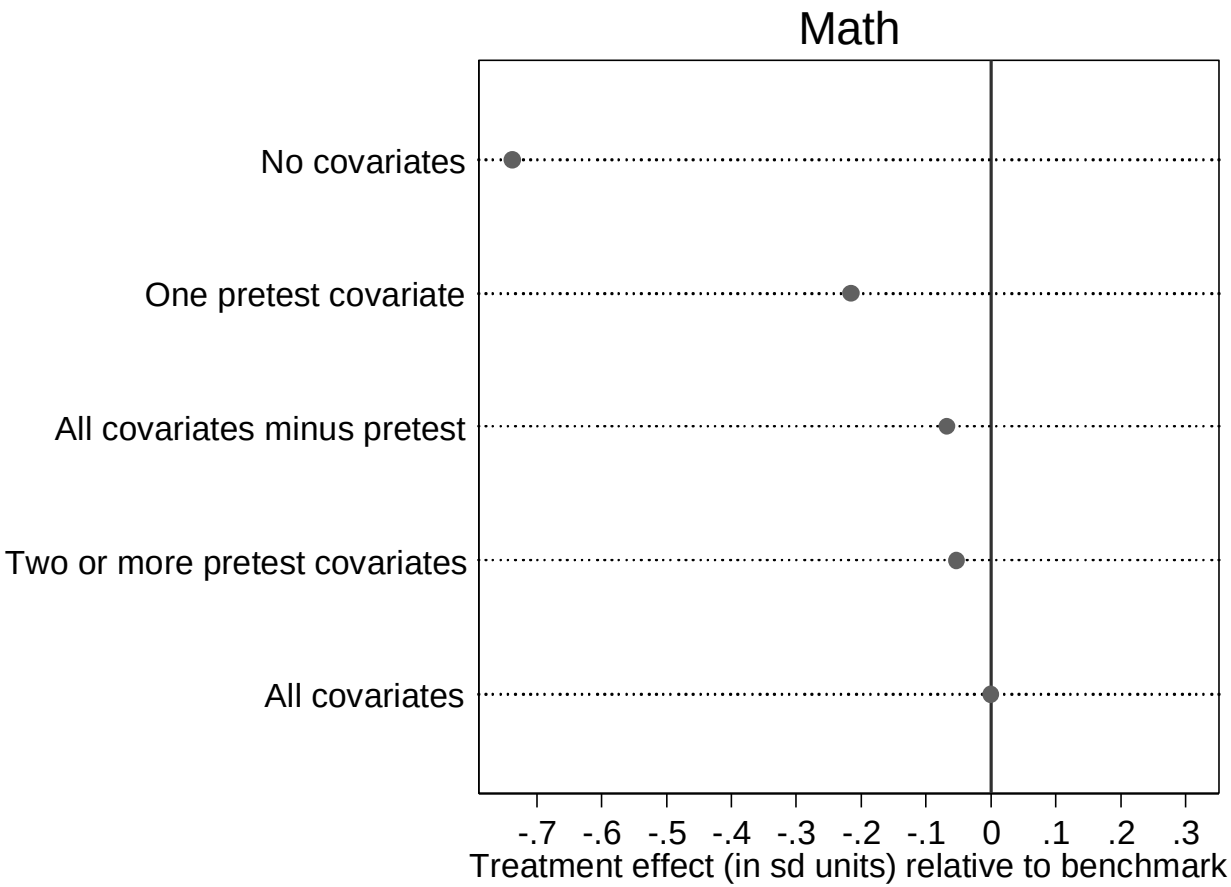
Dataset 1: Correlation with Selection

	Correlation with Retention in Kindergarten	**Correlation Lower Bound**	**Percent of lower bound**
Reading Pretest	-0.185*	-0.38	48.7%
Math Pretest	-0.179*	-0.37	48.4%

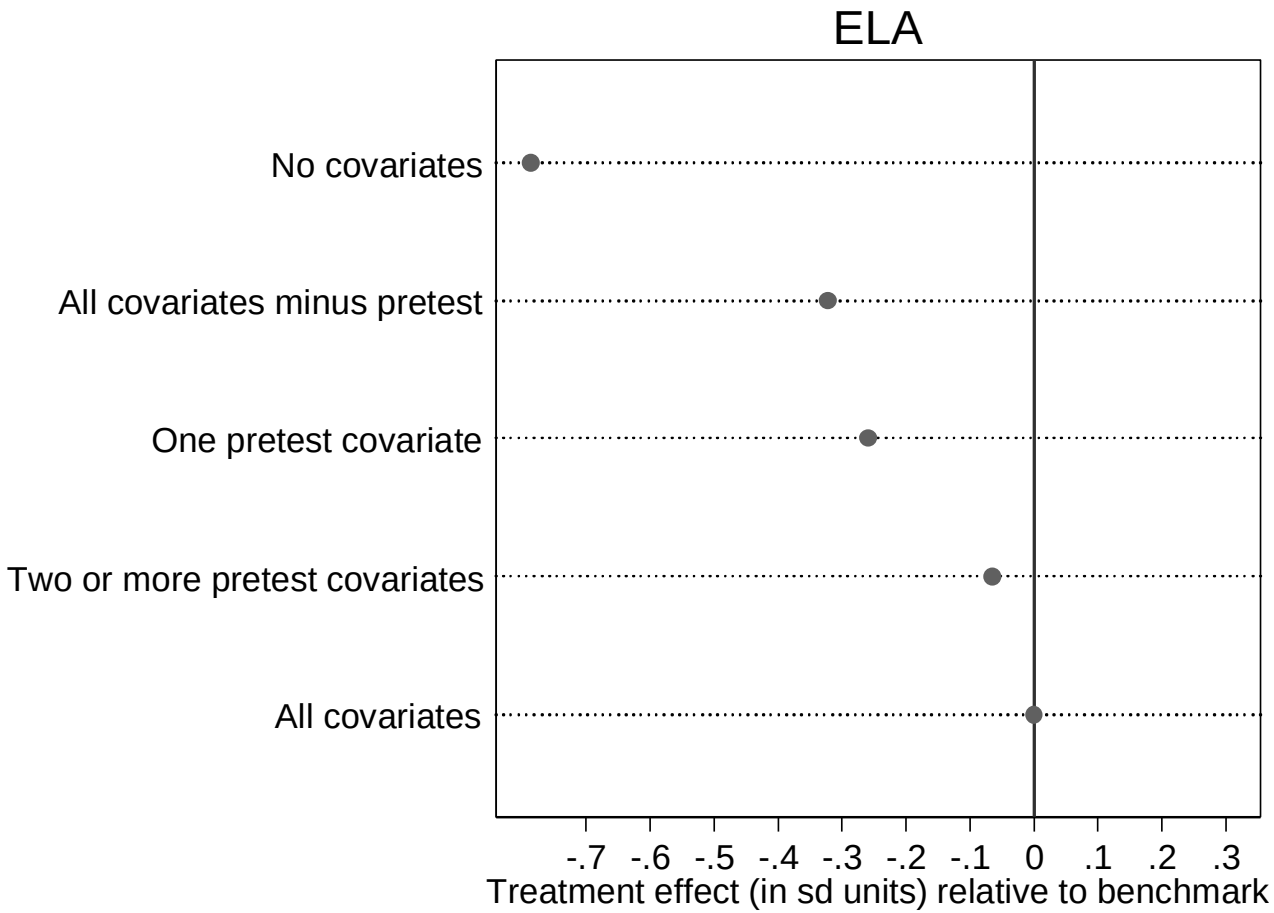
Data set 1: Analytic Approach

- Broke 144 covariates into three groups:
  - One wave of pretest data (spring of K)
  - Two waves (fall and spring of K)
  - 140 other covariates
- Created propensity scores with each cov set and estimated reading and math effects
- Note: Bias reduction compared to benchmark model, not RCT!

Dataset 1: Math Results



Dataset 1: ELA Results



Dataset 2:
Indiana Benchmark Assessment Study (Grade 5)

- 56 K-8 schools 5th graders randomly assigned to:
  - Treatment: state benchmark assess system (n=34)
  - Control schools: business as usual (n=22)
  - Outcomes: Math and ELA ISAT scores
- QE comparison group from all other schools in state serving 5th grade students (n = 681)
- Rich set of student and school covariates with multiple waves of pretest data

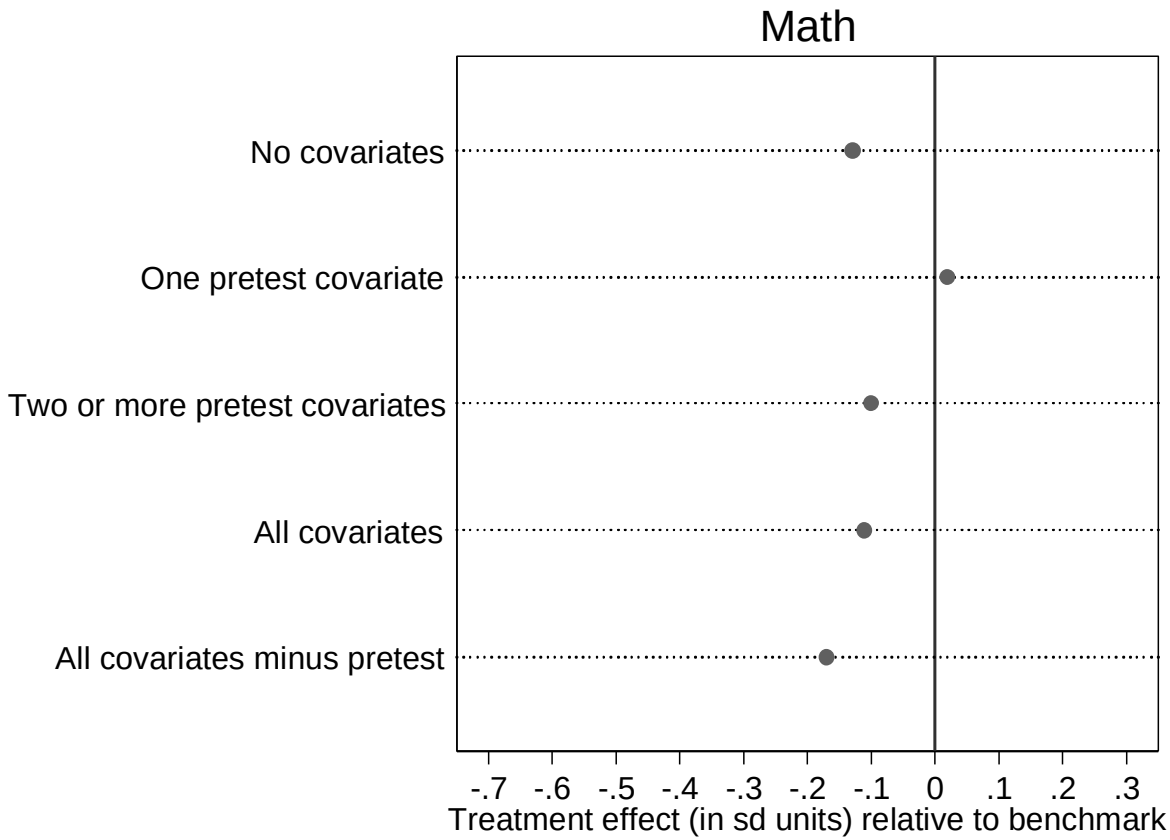
Dataset 2: Selection

- Schools selected into study cos interested in implementing the program
- Principals interviewed and cited
  - Taking advantage of free resource from the state
  - A commitment to data driven decision making
  - Knowledge of other schools implementing
  - No mention of participation due to school's past academic performance – i.e., the pretest

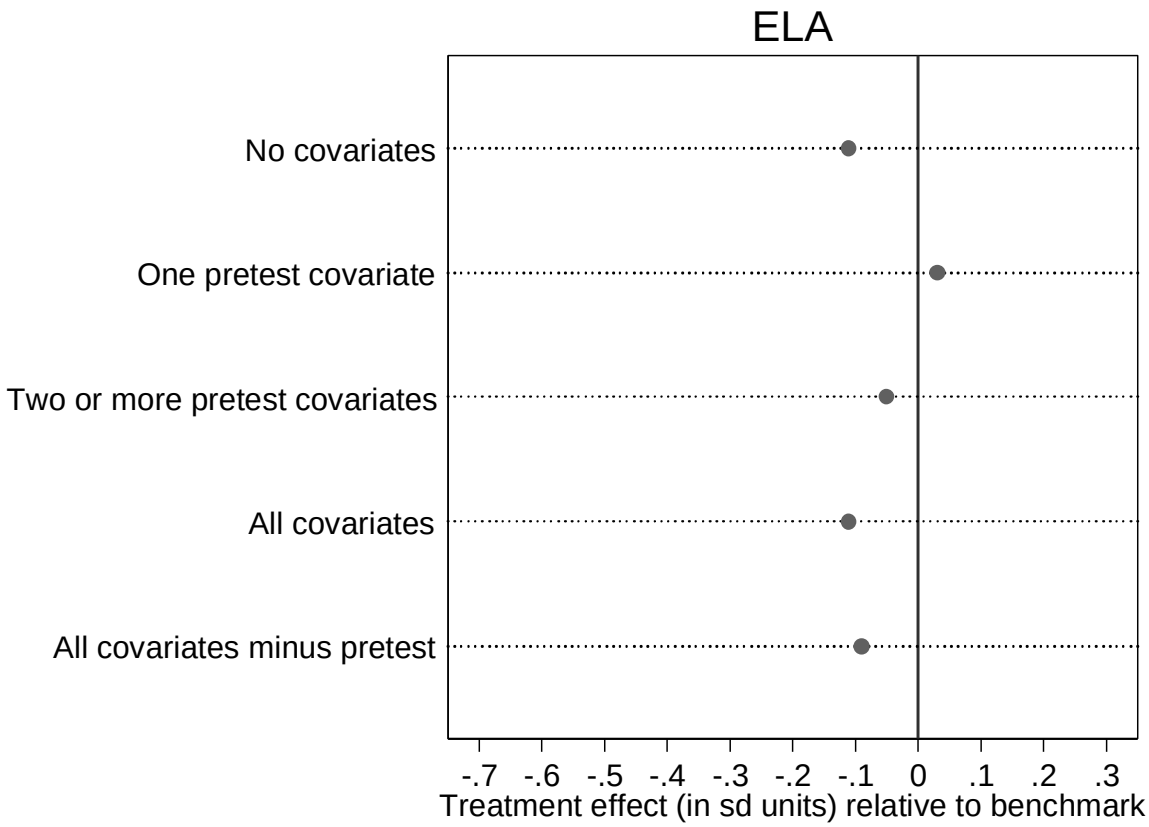
2: No Correlation with Selection

	Correlation with Selection into Benchmark Assessment System
Reading Pretest	0.041
Math Pretest	-0.012

Dataset 2: Math Results



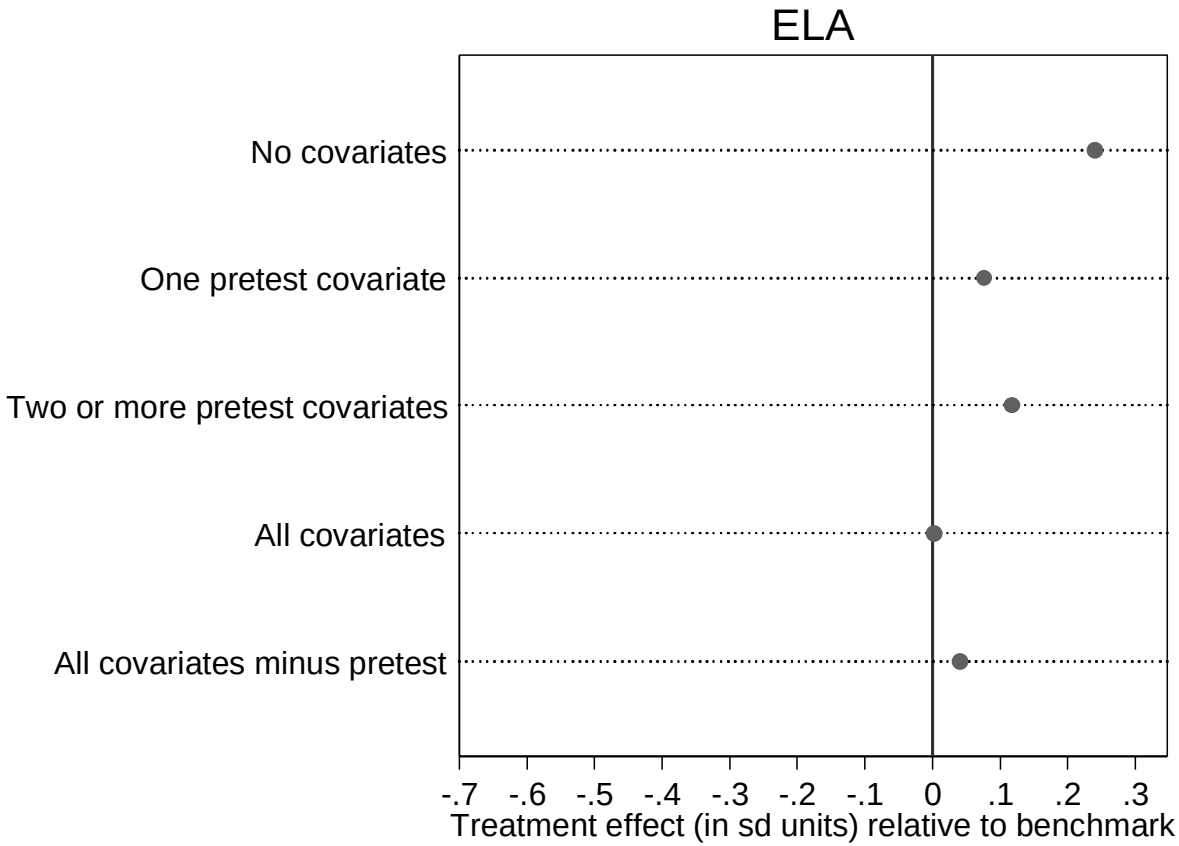
Dataset 2: ELA Results



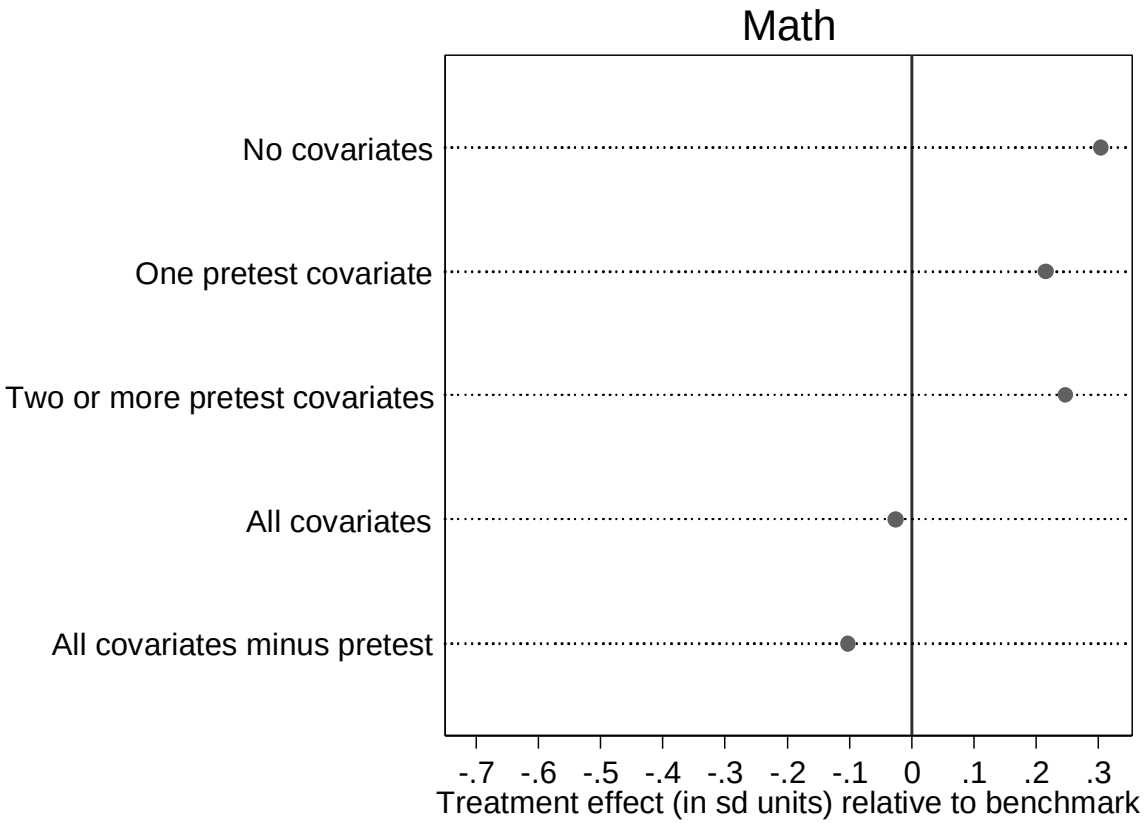
Shadish et al. Correlation with Selection

	Correlation with Selection into Vocabulary Training
Reading Pretest	0.169*
Math Pretest	-0.090

Dataset 3: ELA Results where Pretest and Selection correlate



Math Results where Pretest and Selection not correlate



Summary of Pretest Results

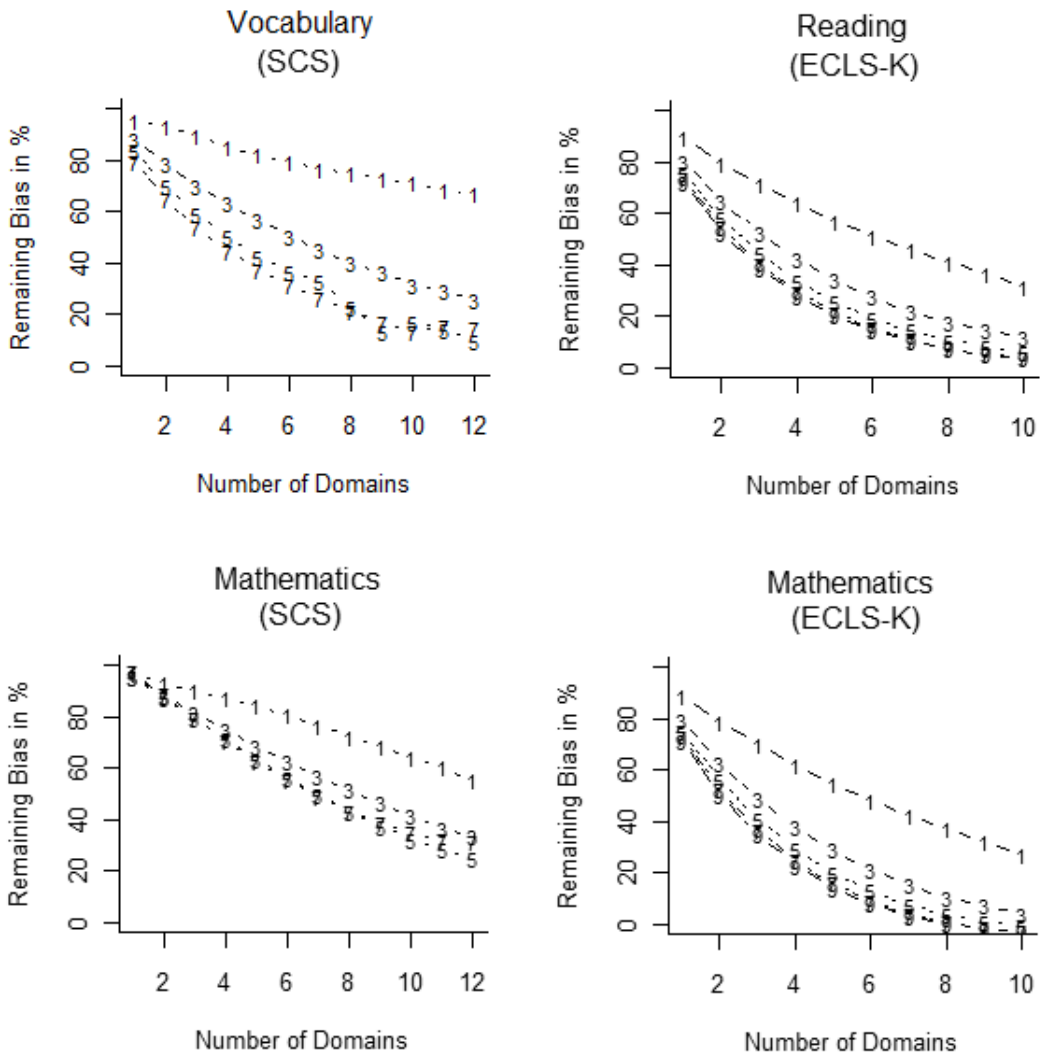
- Cannot assume the pretest is always related to selection, even if it often is
- You should probably always include it
- But yoou are better guided by theoretical explication of all plausible selection processes
- Better supplementing it with more waves and other covariates.

Steiner, Cook & Li (in press)

- What happens if selection unknown
- “Rich” covariates –more constructs and higher reliability
- Theory = pick up of true but unknown selection process
- Two data sets – one with 156 covariates at one pretest and the other with 144 over two pretest time points.
- Each has reasonable theory of selection; we identify it and then throw it away to ask: How do the remaining covariates function collectively if individually not good

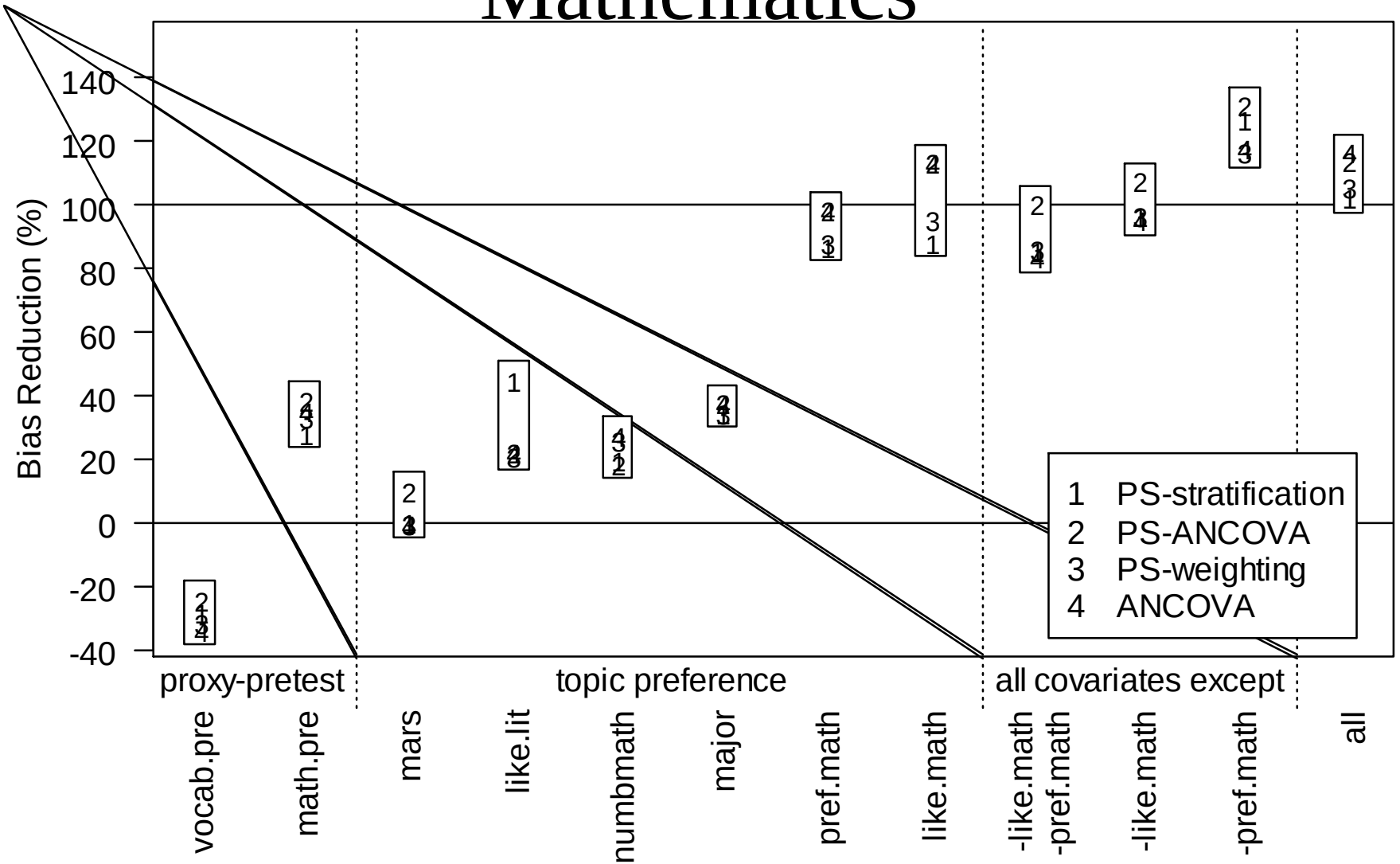
With all Covariates

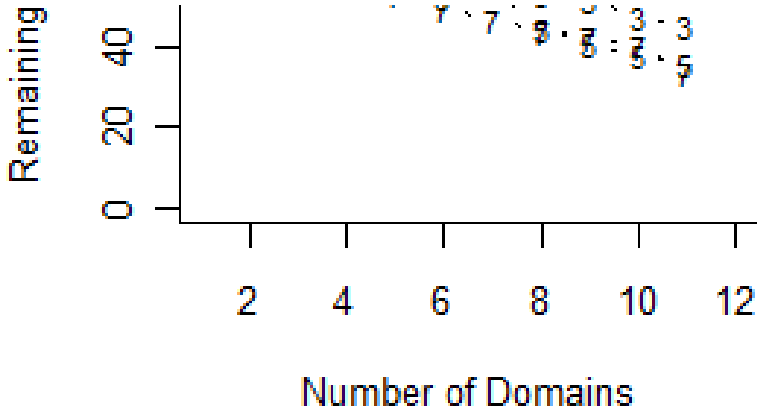
- Partitioned into Number of constructs
- Into number of items per Construct
- Question is: How is bias reduction affected by number of constructs and their reliability under 2 conditions:
- A. When all covariates are there
- B. When effective single covariates removed

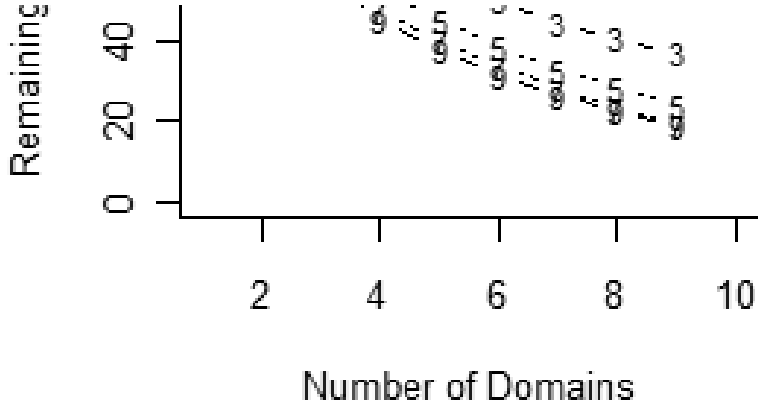

Vocabulary
(SCS)
Reading
(ECLS-K)
Mathematics
(SCS)
Mathematics
(ECLS-K)
Remaining Bias in %
Number of Domains
0
20
40
60
80
2
4
6
8
10
12

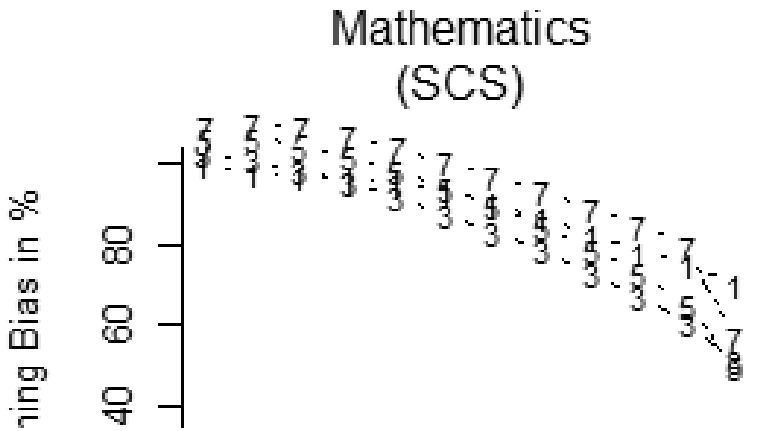
Remove effective single covariates

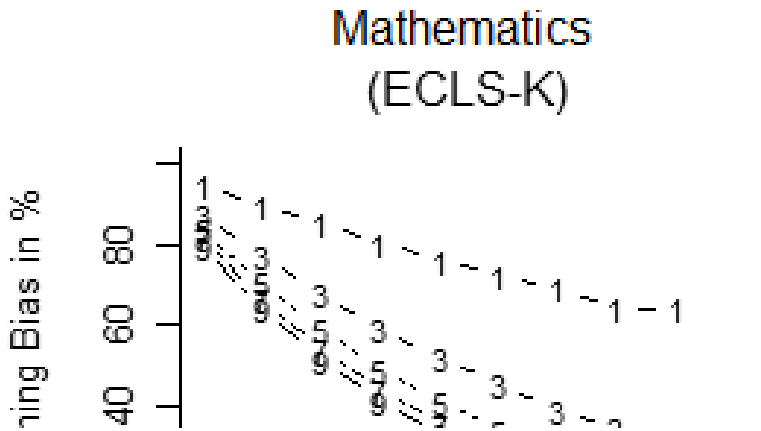
Mathematics




Remaining
0
20
40
2
4
6
8
10
12
Number of Domains


Remaining
0
20
40
2
4
6
8
10
Number of Domains


Mathematics
(SCS)
ning Bias in %
40
60
80


Mathematics
(ECLS-K)
ning Bias in %
40
60
80

If Bigger Data = More Local Comparisons

- Big effect on bias reduction in general, but not always effective
- Combining it with criteria for adequate match and focal matching where the local match is not adequate may well work better
- But focal matching not so easy (as we will see)

If Bigger Data = More Time Series

- When linked to comparison groups also,
- CITS in 11 of 11 cases has done good job
- But matching may be analytically more robust than modeling baseline time trends

If Bigger Data = More Reliability

- This will definitely help
- Long known that unreliability in covariates a source of bias

If Bigger Data = More pretest availability

- Pretest best single variable in general for reducing bias
- But not always the case
- When pretest is corr with selection does a good job of reducing selection
- When pretest not corr with selection, in Ed. often little selection to account for. Is most selection via pretest or correlates thereof – at least for achievement

If Bigger Data = More Constructs assessed at Pretest Time

- Will remove most bias if selection completely or largely known and well measured
- Will remove some bias if selection not well known but “rich” covariate set. May not be enough
- Will remove some but not most bias if ad hoc set of measures
- Some measures never help -- demographics

Bottom Line

- Bigger data will not increase RA except for low profile experiments
- Will improve QEs in general by mechanisms discussed
- Never be as good as RCTs
- But will they be good enough in terms of other attributes in decision theory?
- What has happened to survey research…

Data-Bound Summary

- Although no meta-analysis, things look good for RD, CRD, and CITS
- Looks very good when you design prospective studies and include measures to account for multiple possible selection processes
- Intact, focal and local matching each sometimes reduce all bias, almost always reduce some bias, but likely best together in hybrid matching
- It is clear that pretests do not always reduce bias, but the smart money is that they will sometimes reduce all bias and that they will often be a significant part of a bias reduction strategy with other sampling and covariate choice models (
- We anticipate this presentation would be very different five years from now, not so much with respect to RDD and ITS, but with respect to non-equivalent control group designs.

Broader Summary

- Let us all acknowledge that RCT is best in theory and not get into meaningless fights.
- Let's ask: is the assumption warranted that the RCT is "far" superior for warranting causal inference?
- Is an evidence-based empirical rationale already emerging for including some QE studies as *acceptable* contributions to evidence-based policy suggestions?

Broader Summary

- The second assumption is that evidence-based policy will be better if we have more info about external validity so as to learn about robustness or conditions under which effect sizes vary for the same treatment and effect
- Will having more acceptable studies in our knowledge compendia promote external validity, an Achilles Heel of much evidence-based practice research?

END and THANKS