What are better quasi-experimental practices for evaluating education programs? Comparison with experimental results

Thomas D. Cook, Northwestern University
Kelly Hallberg, American Institutes for Research

Funded by NSF Grant DRL-1228866

Purposes

- We assume RCTs are best way of identifying causal relationships, but that such experiments are not always possible
- Demonstrate some quasi-experimental practices reproduce essentially the same results as experiments
- Theoretically trivial to show this can happen, practically useful to show *when* this can happen

Not Our Purposes

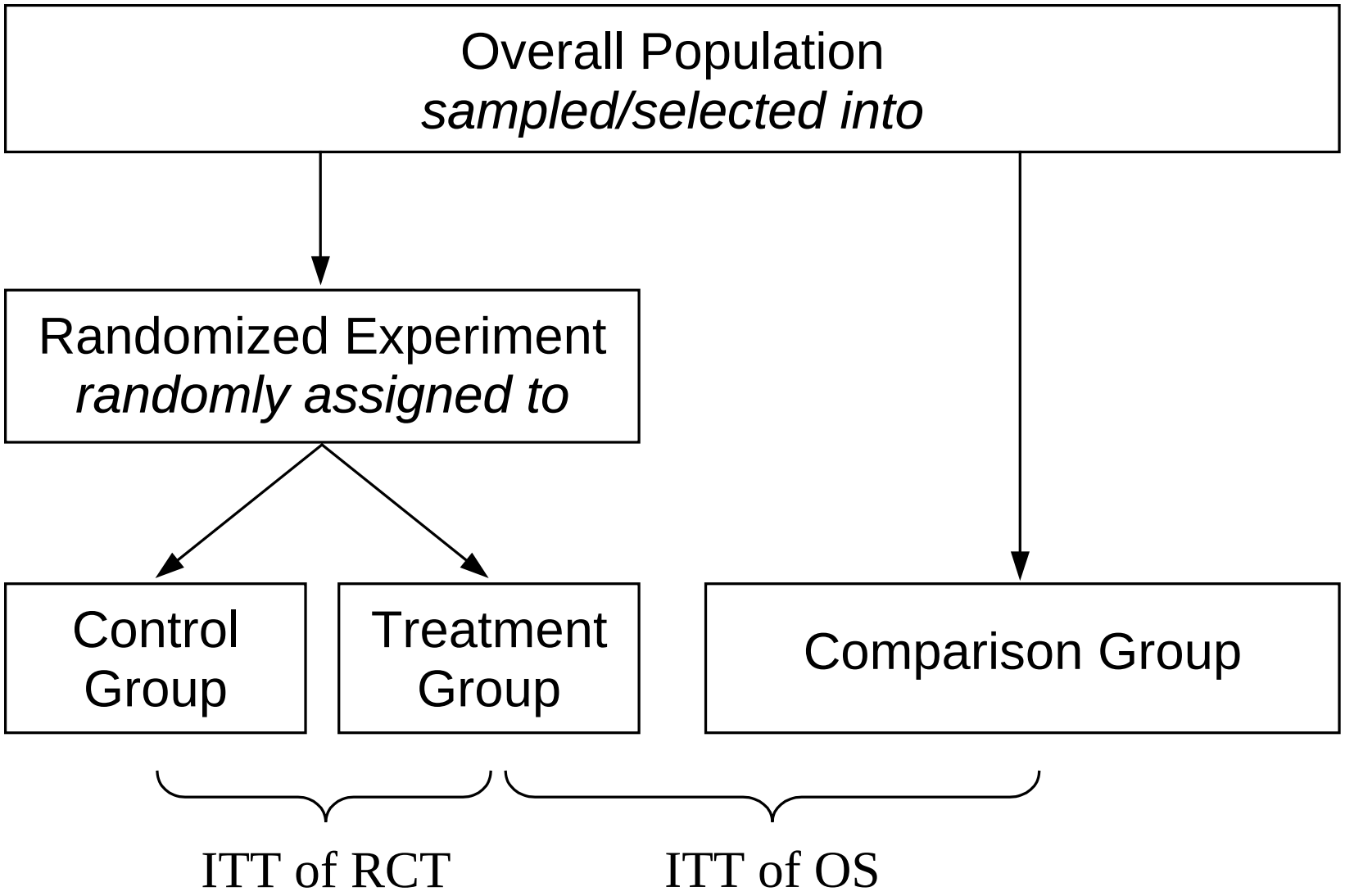
- To argue that even the best quasi-experimental practices are better than experiments for internal validity
- To show how you can take systematic advantage of longitudinal survey data
- A primer on statistical analysis of quasi-experiments. Concentrate more on control via study sampling design, measurement plan and experimental design

How do you know the RCT and QED Results the same: What is a Within Study Comparison?

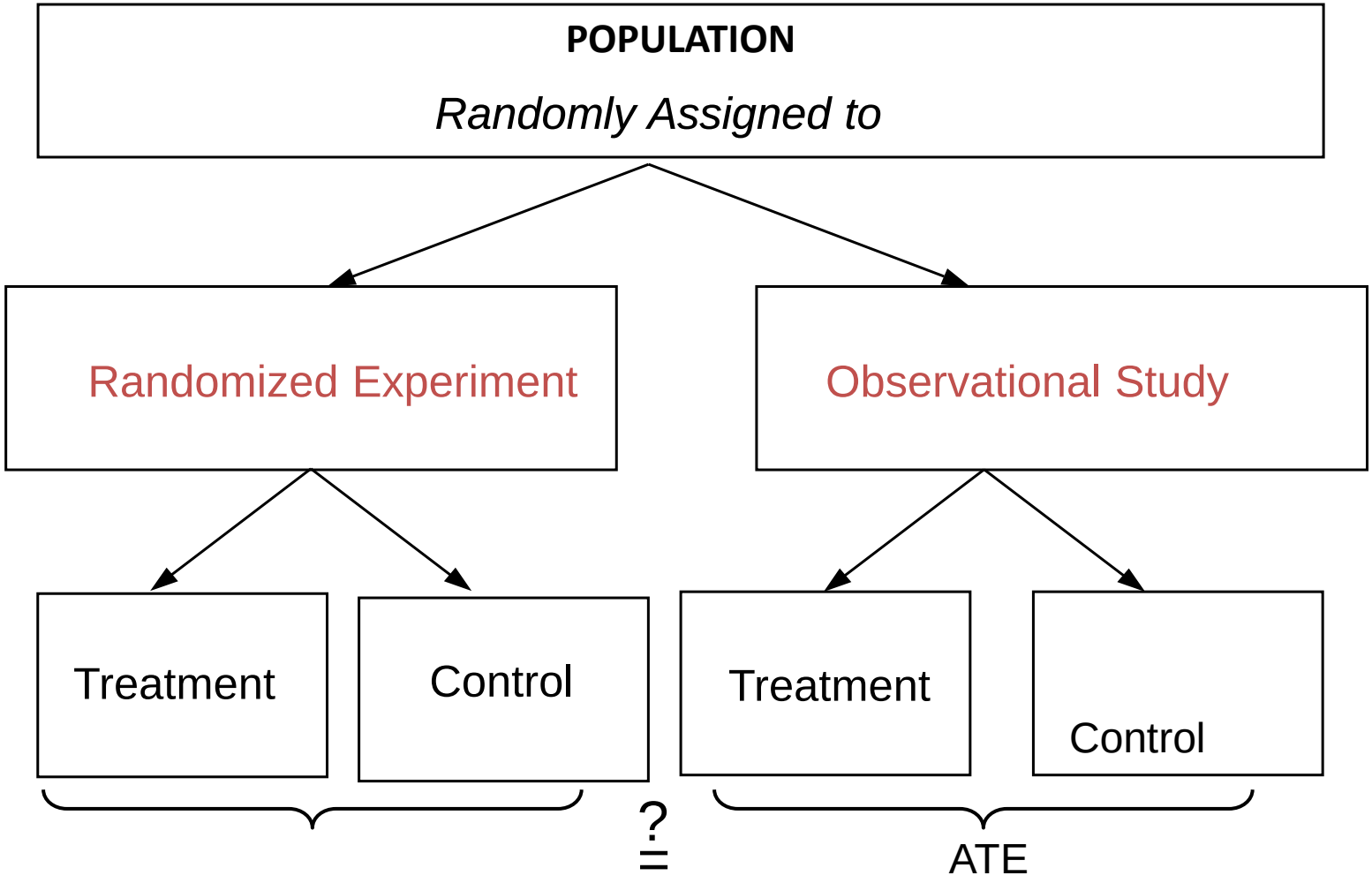
- Examines the comparability of results of RCTs and QEDs

- Investigates the consequences of assigning treatments at random versus systematically without confounds from other sources

WSC: Three-Arm Design



WSC: Four-Arm Design



Conditions for a good WSC

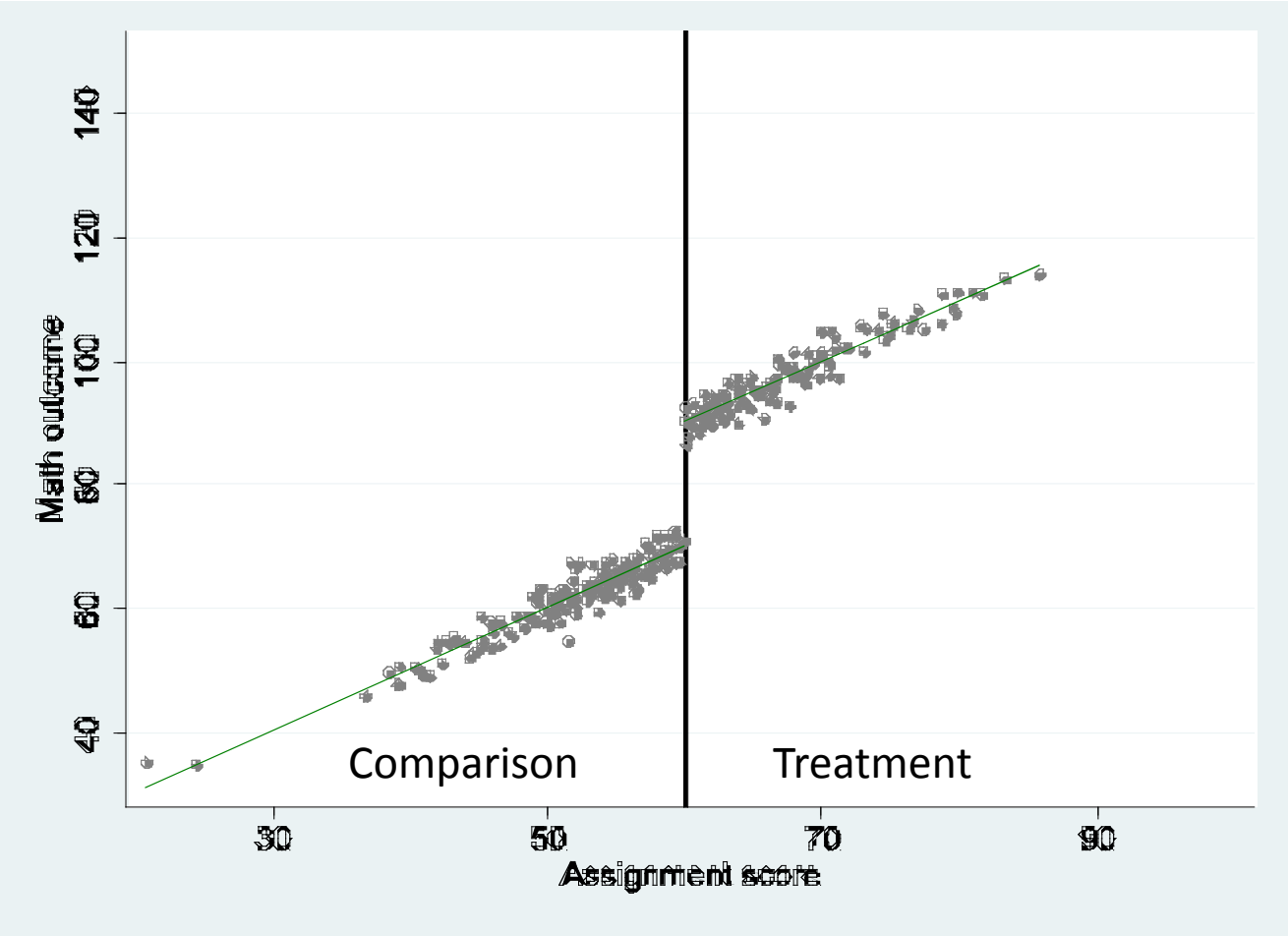
- A well implemented RCT nec for gold standard
- No third variable confounds
- Comparable estimands
- Blinding to the RCT or adjusted QED results
- Defensible criterion for correspondence. Not use similar point estimates, or stat sig patterns or policy recommendations. Instead we will ask how close are the RCT and QED estimates.

Limitations of WSCs

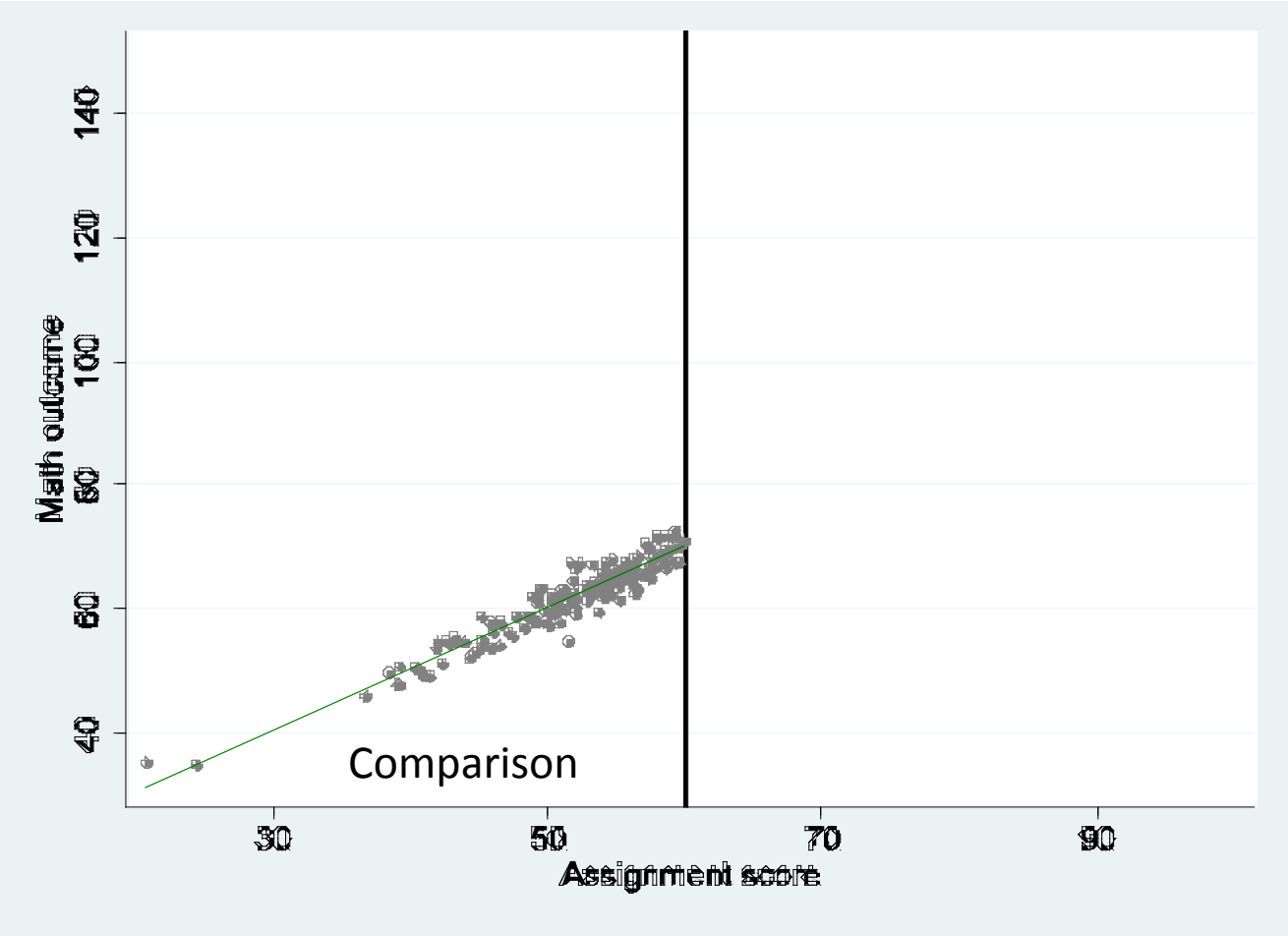
- WSCs can only be done on topics where an RCT is possible
- No reason to believe that a given QED design will always replicate experimental findings
- Seek to identify designs more likely to replicate findings. This requires a large sample size of WSCs for induction
- More definitive conclusions requires more WSCs than we have today, especially for NEQCD – less so for RD and ITS

COMPARING RCT TO RD: WHAT IS A REGRESSION-DISCONTINUITY DESIGN

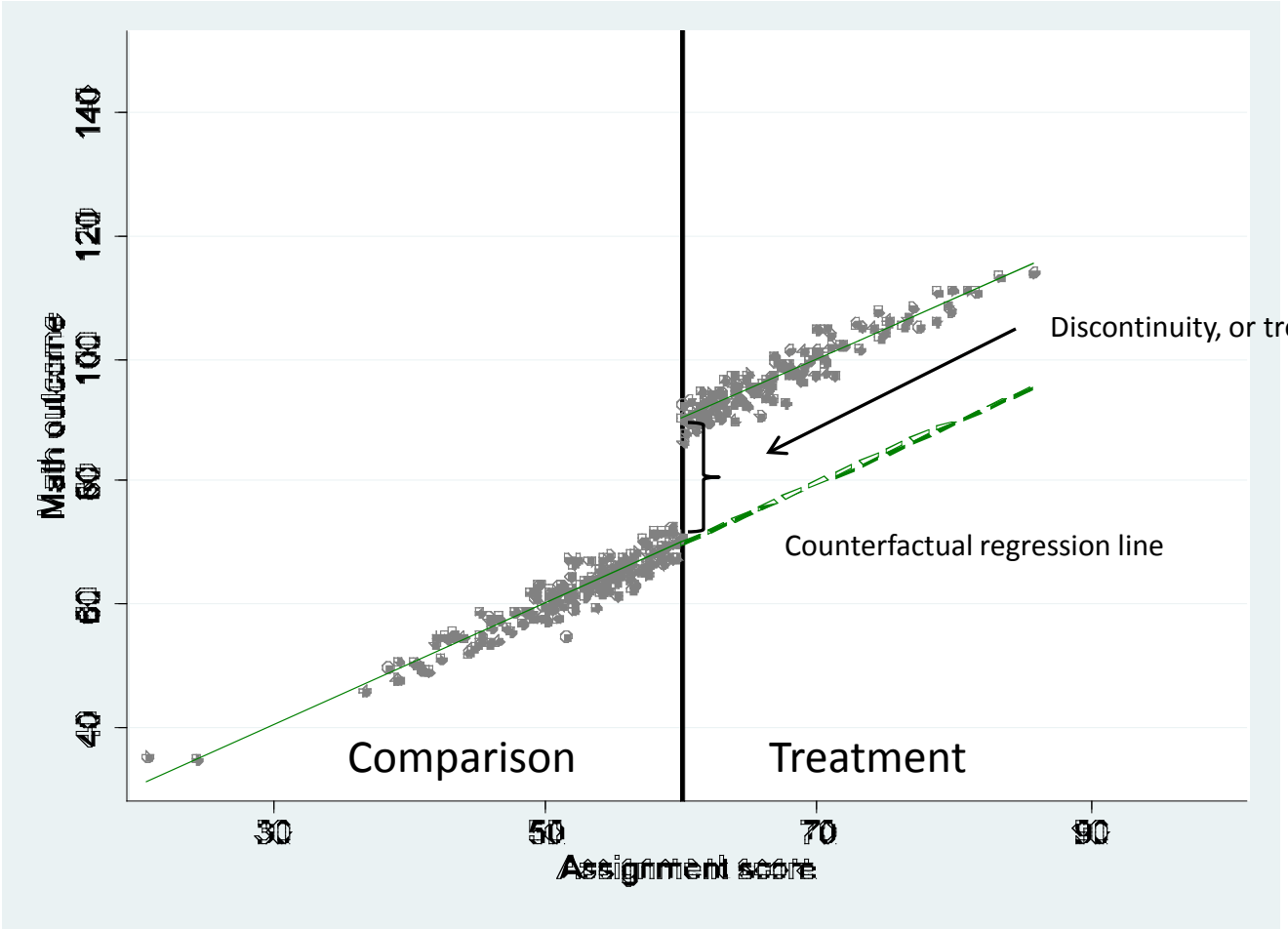
RDD Visual Depiction



RDD Visual Depiction



RDD Visual Depiction



Limitations of RDD

- Functional form assumptions
- Causal generalization
- Statistical power relative to RCTs
- Treatment crossovers and manipulation

WSC Results for RDD

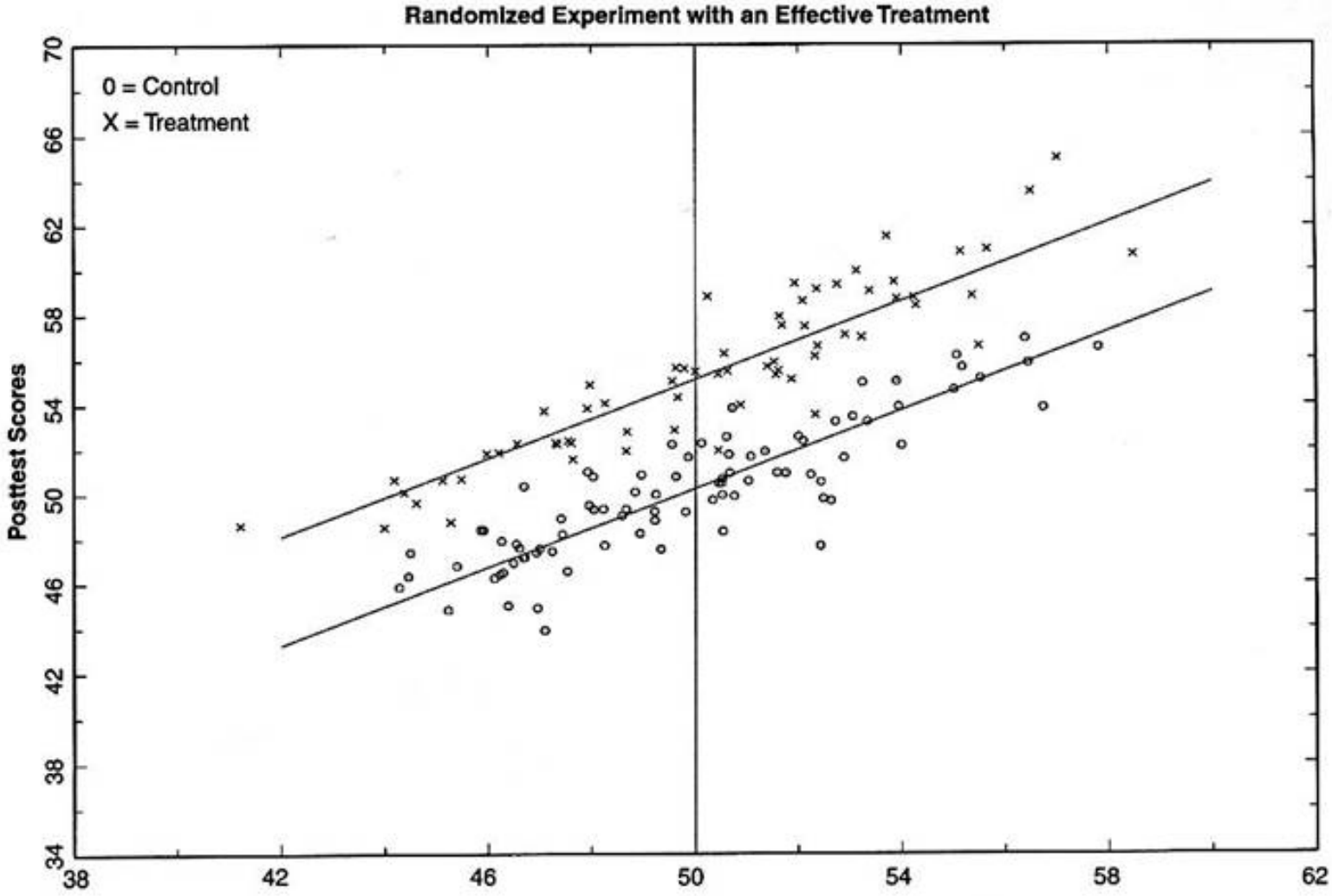
- There have now been 8 WSCs and all report the same finding – comparable effect estimates at the cutoff
- No formal meta-analysis yet
- Theoretically trivial, but does demonstrate that the implementation of both RCTs and RDDs are generally good enough to get the same results in real practice

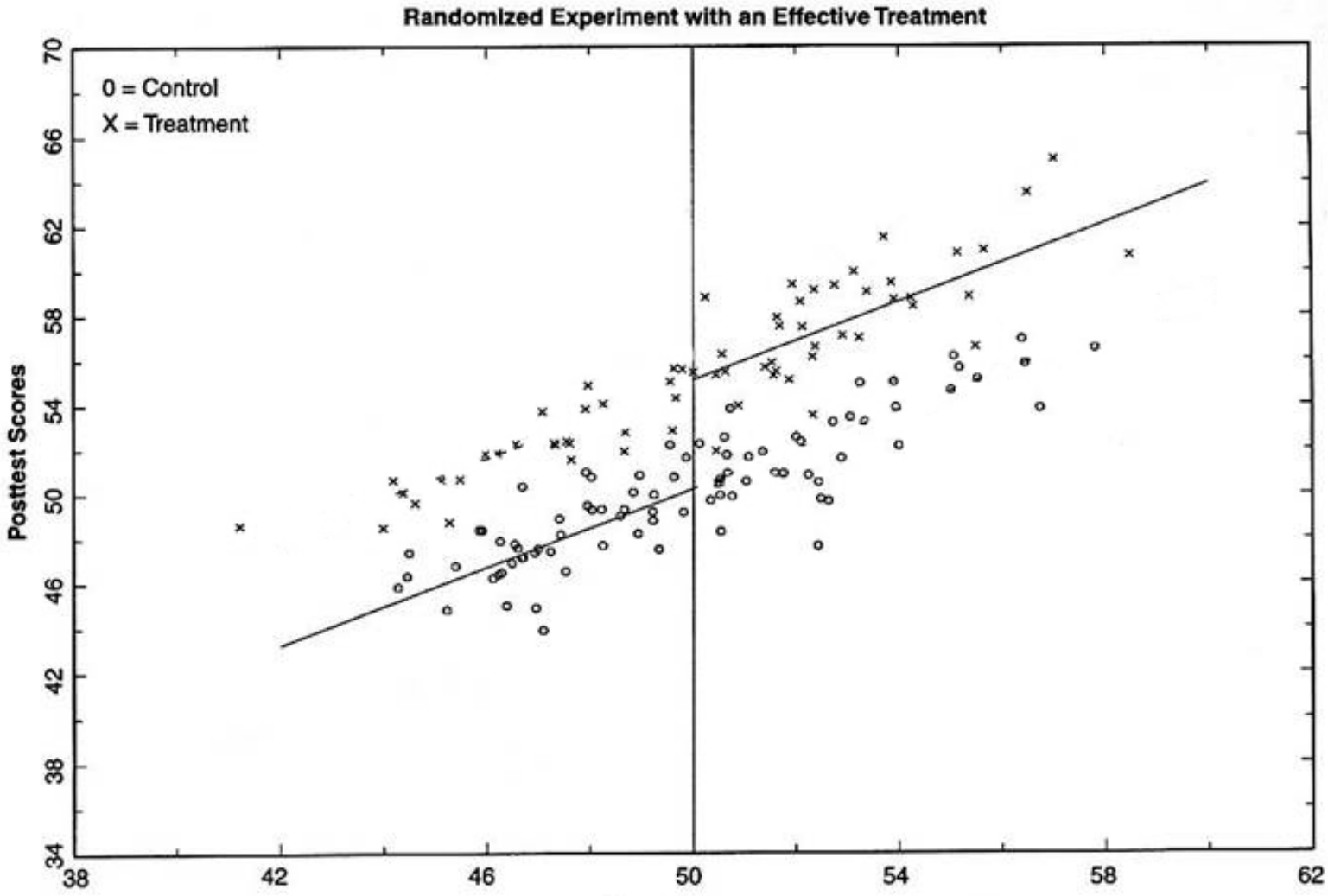
Overcoming the limitations of RDD:

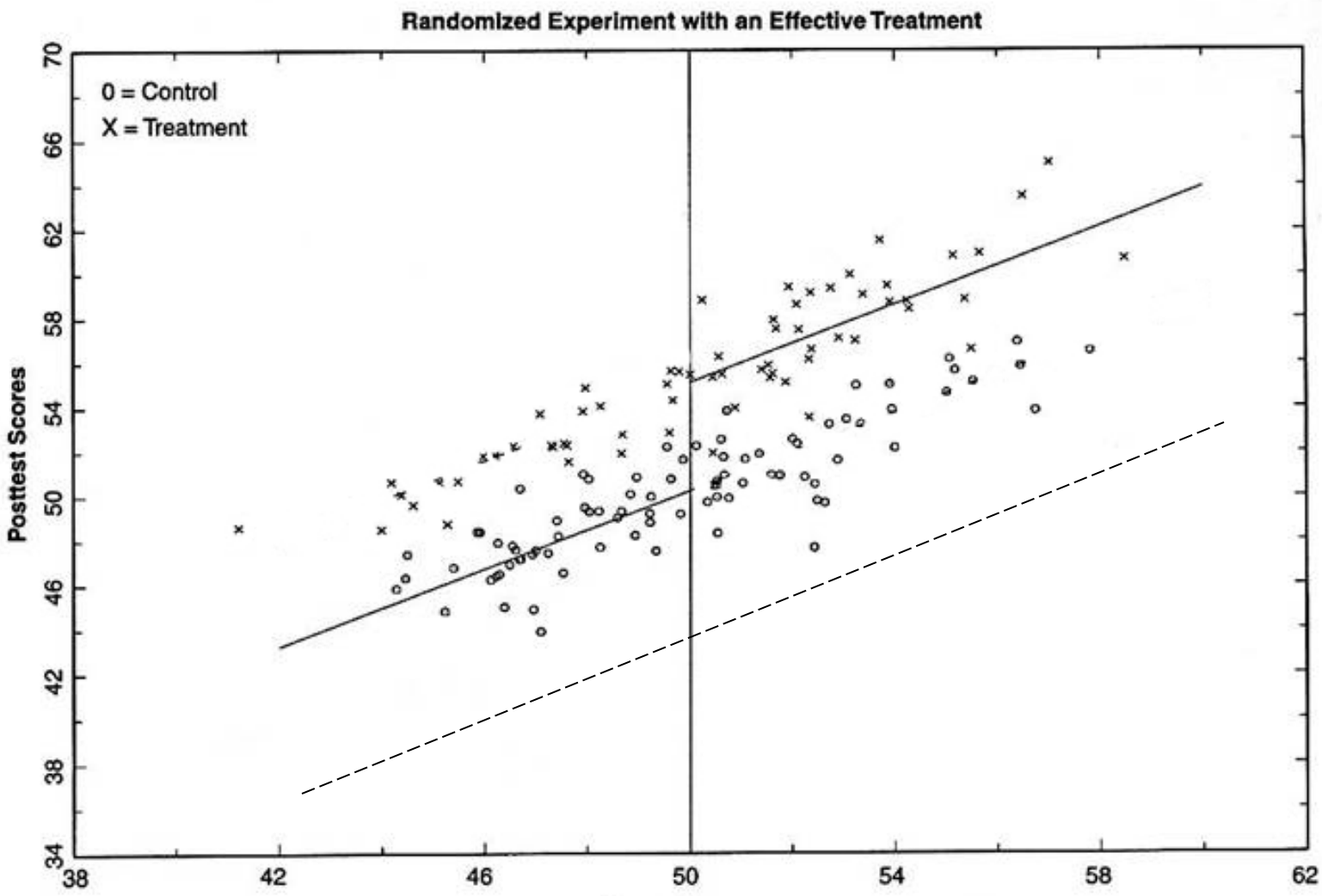
- Adding a pretest RD function is currently recommended for RCTs too
- How does this help with functional form estimation in RDD?
- Causal generalization?
- Statistical power?

Wing & Cook (2013)

- The RCT involved: having control over the budget for services in disabled families vs. the families determining the services with the same resources available
- Dependent variable: amount actually spent for disability services
- How to create the CRD from an RCT?


Randomized Experiment with an Effective Treatment
0 = Control
X = Treatment
Posttest Scores
70
66
62
58
54
50
46
42
38
34
38
42
46
50
54
58
62


Randomized Experiment with an Effective Treatment
0 = Control
X = Treatment
Posttest Scores
70
66
62
58
54
50
46
42
38
34
38
42
46
50
54
58
62


Randomized Experiment with an Effective Treatment
0 = Control
X = Treatment
Posttest Scores
70
66
62
58
54
50
46
42
38
34
38
42
46
50
54
58
62
Posttest regression
Pretest regression

Comparative RD Design

- Comparison here = payments made before the RCT began (pretest measure of the outcome
- How it improves functional form estimation
- How it improves power
- How it facilitates causal generalization away from the cutoff point

Comparative RD Results in Standardized Difference from RCT away from Cutoff: Non-Parametric Analysis

State	Cutoff Age = 35	50	70
AK	.07	.05	04
NJ	.19	.13	.12
FL	-.09	-.08	-.04

Summary for RDD

- No meta-analysis yet, but little doubt that unbiased causal inference results at the cutoff point in studies that people actually do
- There is some empirical reason to believe that the unique generic limitations of RDD can be mitigated or are mitigated by adding a reliable RDD function from at least pretest data
- We should call for more comparative RDD designs in the future
- Three other replications are underway with educational data

COMPARISON OF ITS AND RCT

Interrupted Time Series Can Provide Strong Evidence for Causal Effects

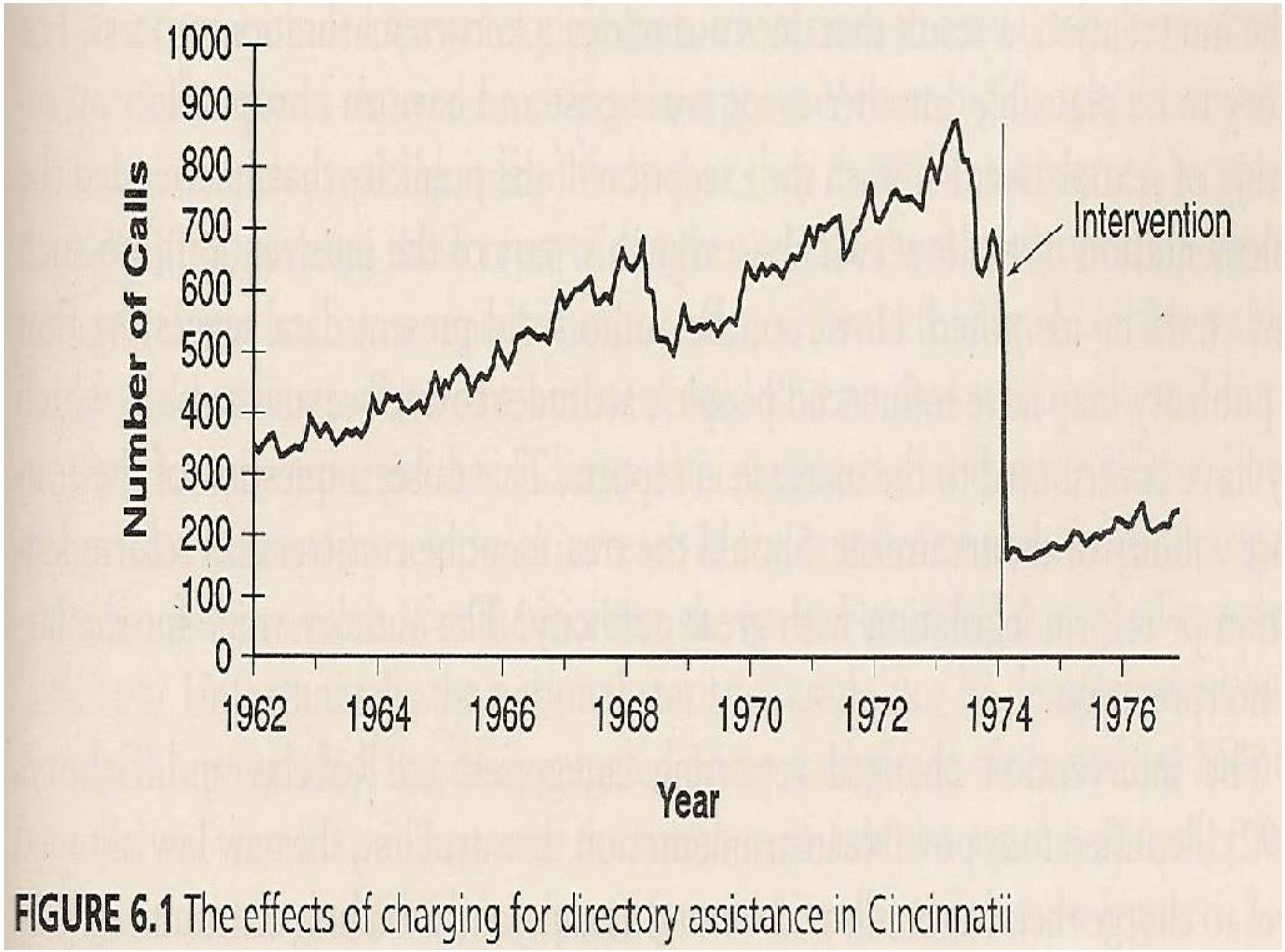
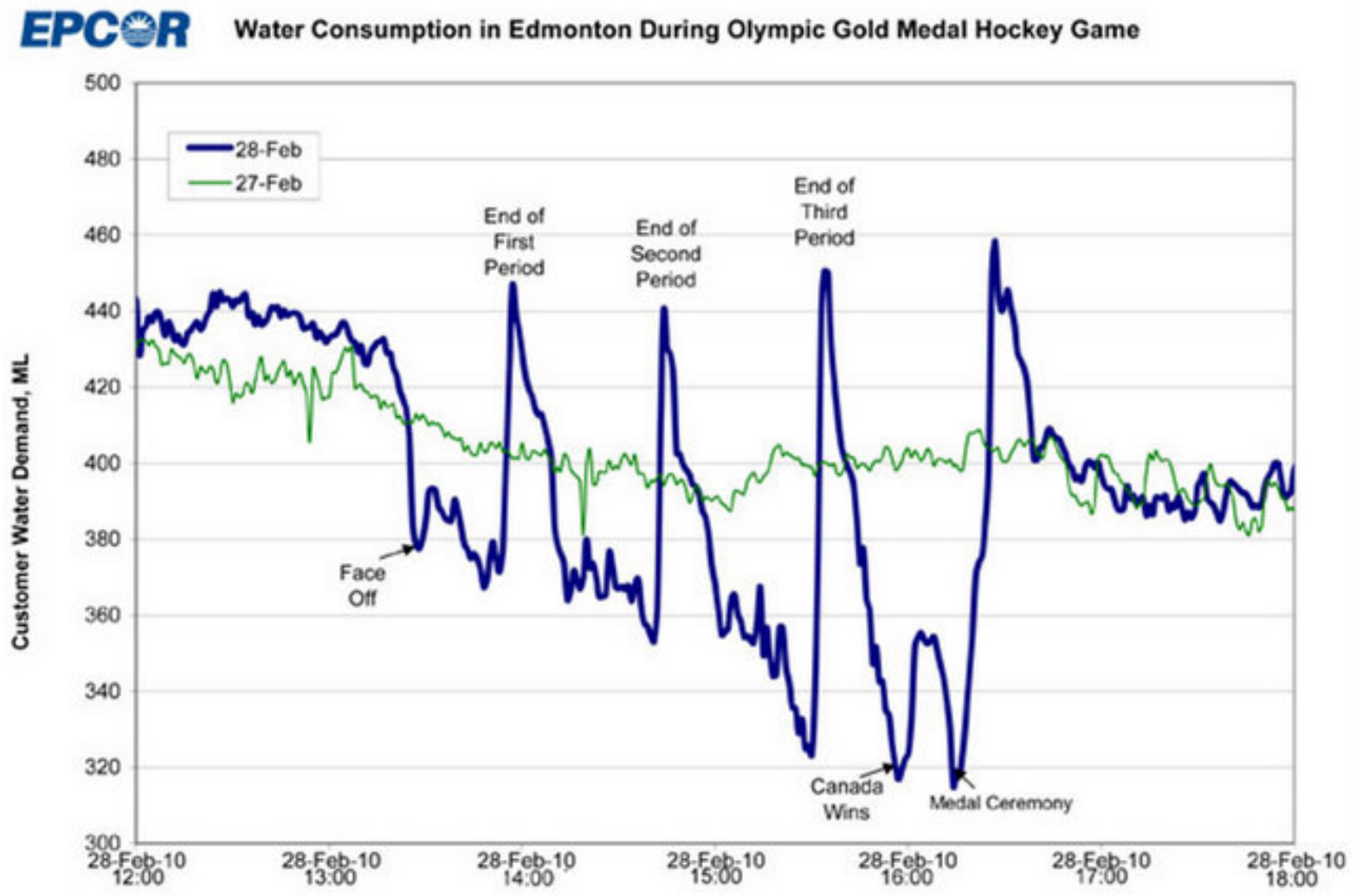


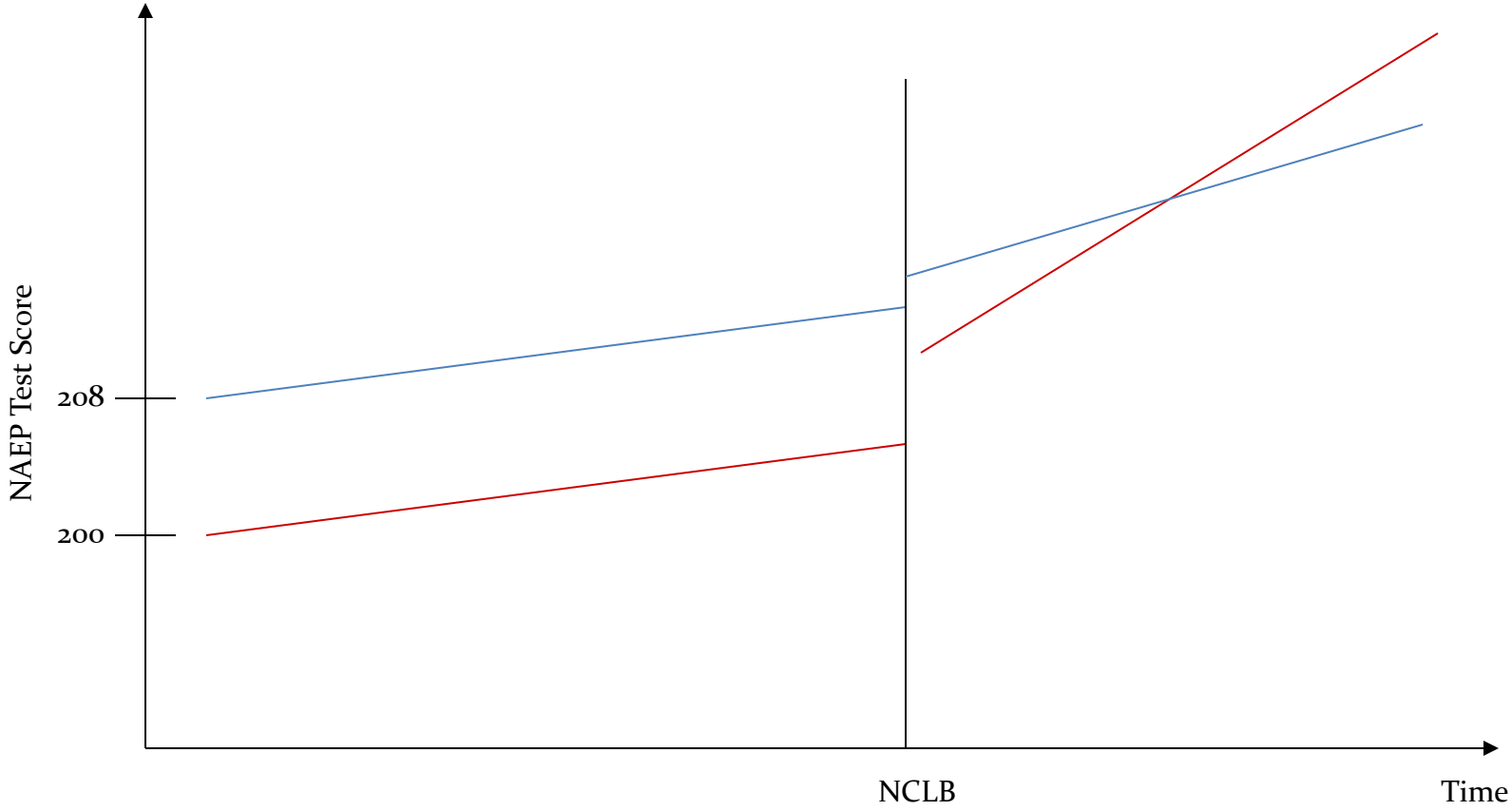
FIGURE 6.1 The effects of charging for directory assistance in Cincinnatii

- Clear Intervention Time Point
- Huge and Immediate Effect
- Clear Pretest Functional Form + many Observations
- No alternative Can Explain the Change

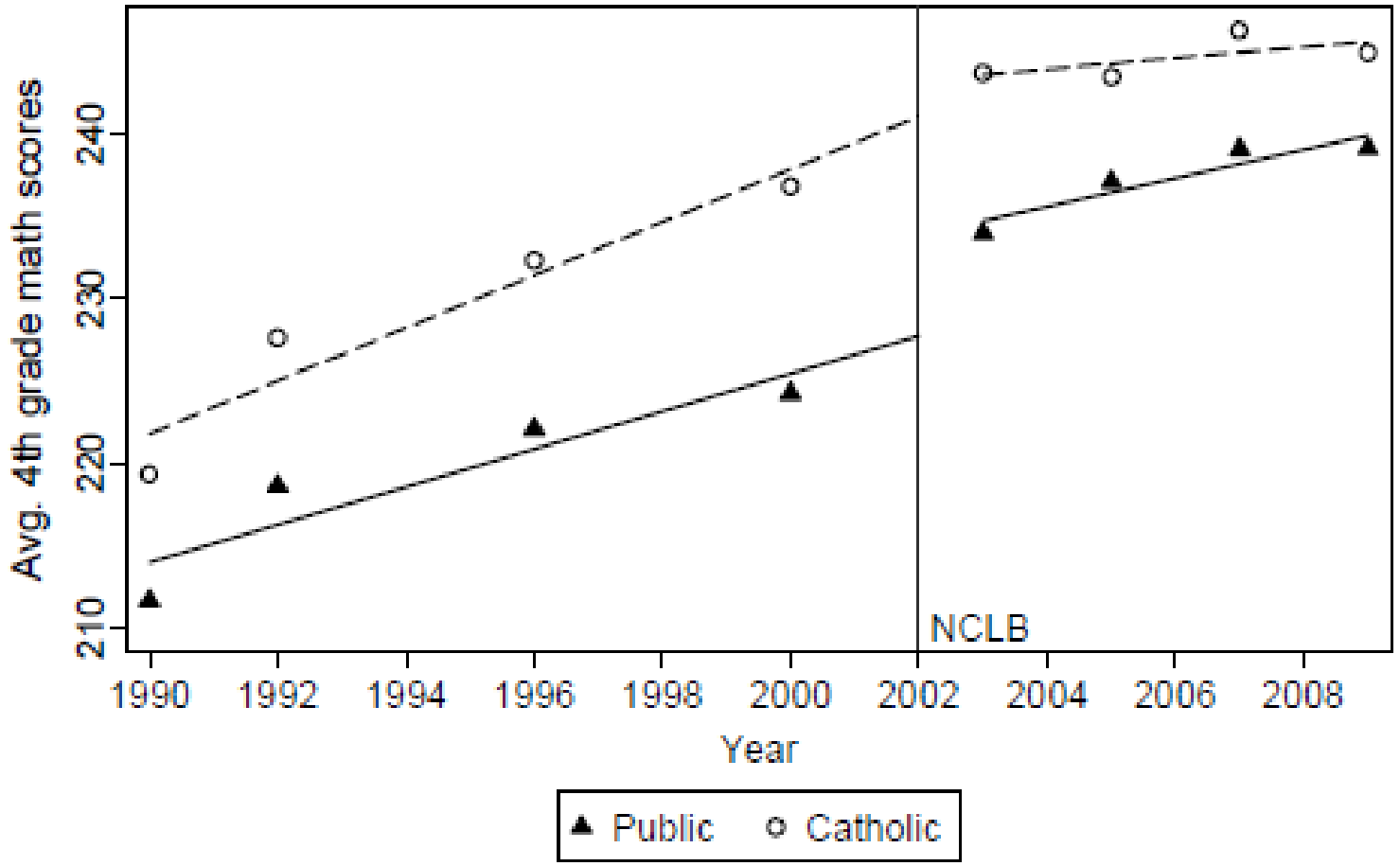
What if Everyone in Canada Flushed at the Same Time



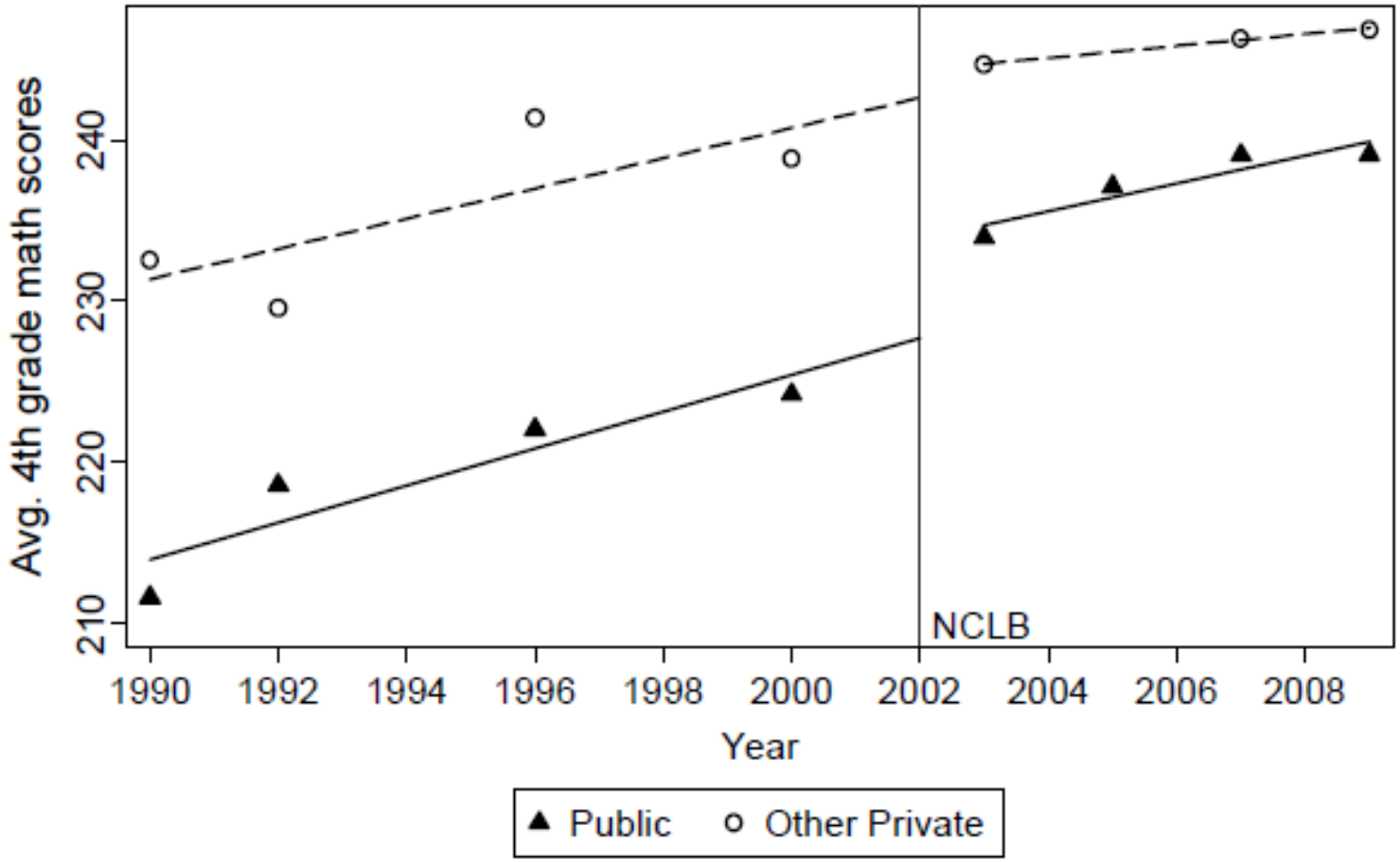
Hypothetical NCLB effects on public (red) versus private schools (blue)



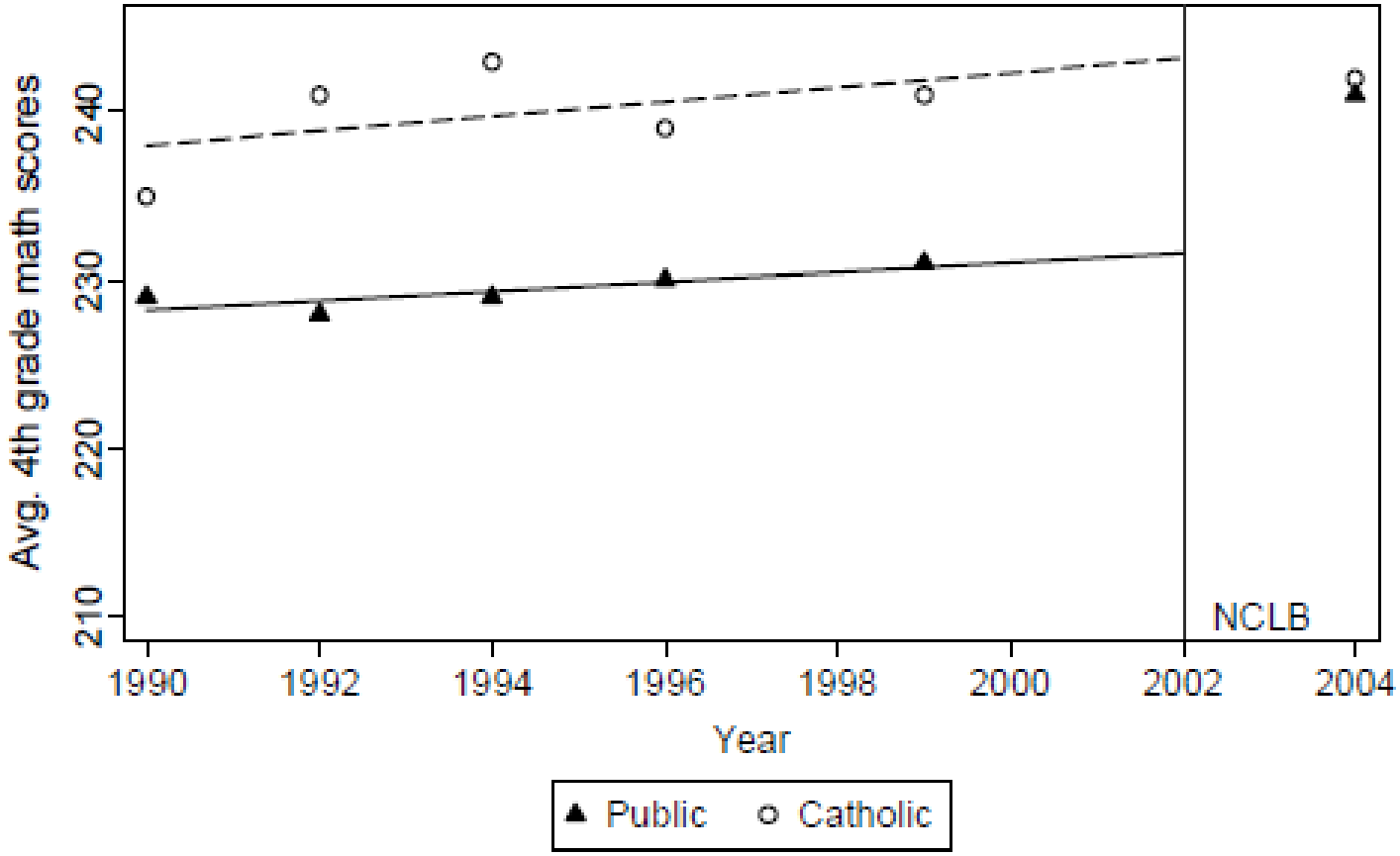
Main NAEP 4th grade math scores by year: Public and Catholic schools



Main NAEP 4th grade math scores by year: Public and Other Private schools



Trend NAEP 4th grade math scores by year: Public and Catholic schools



Limitations of ITS

- History alternative explanations around the intervention point
- Functional form extrapolation needed
- Analysis has to account for correlated errors (we will not deal with this issue here)
- First two points, suggest the advisability of a comparative ITS

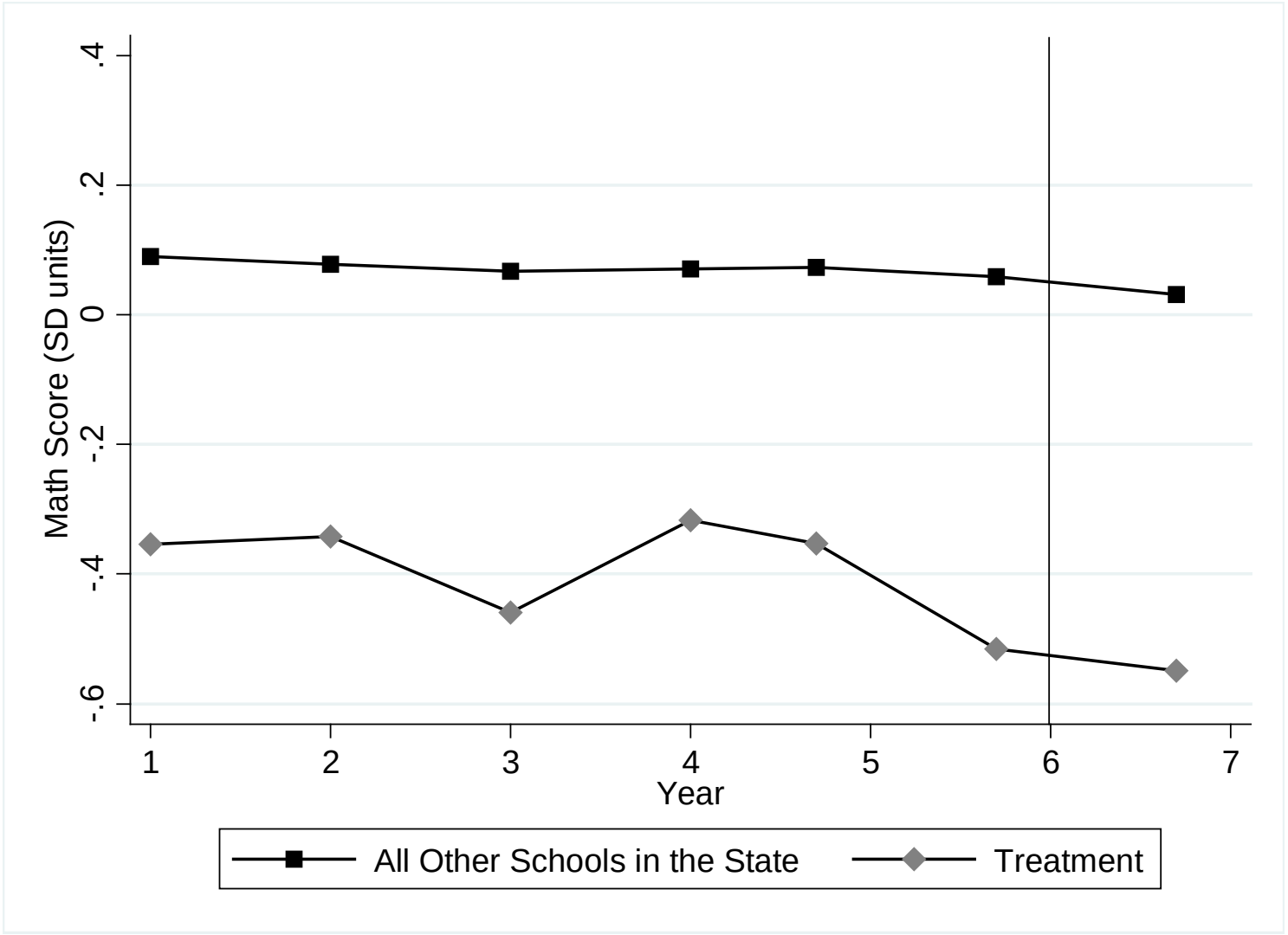
WSC and ITS

- Four studies in medicine, two in education
- All claim not dissimilar causal inferences
- No meta-analysis to date
- No analysis of file drawer problem

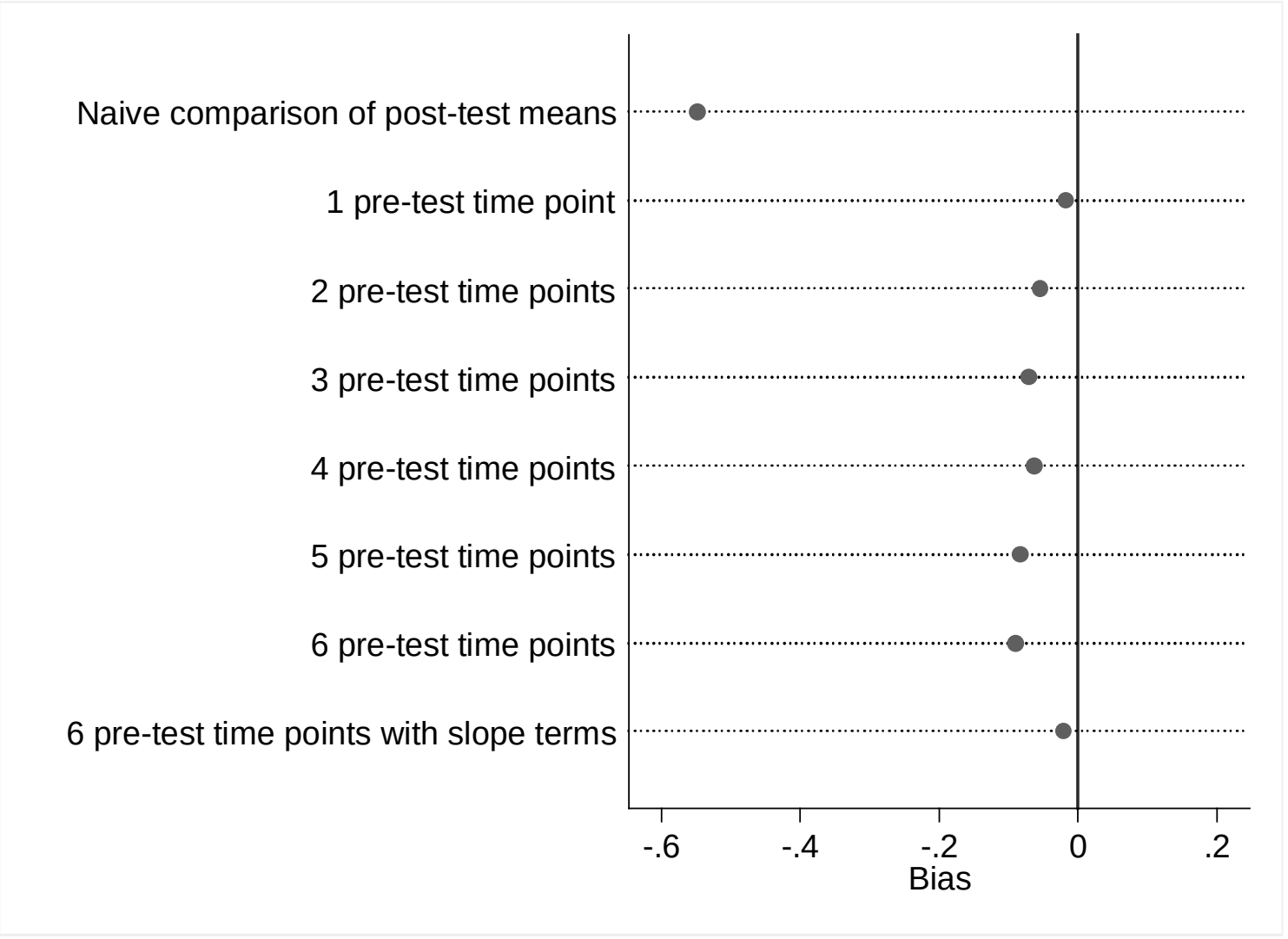
St. Clair, Cook, & Hallberg (In Press)

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

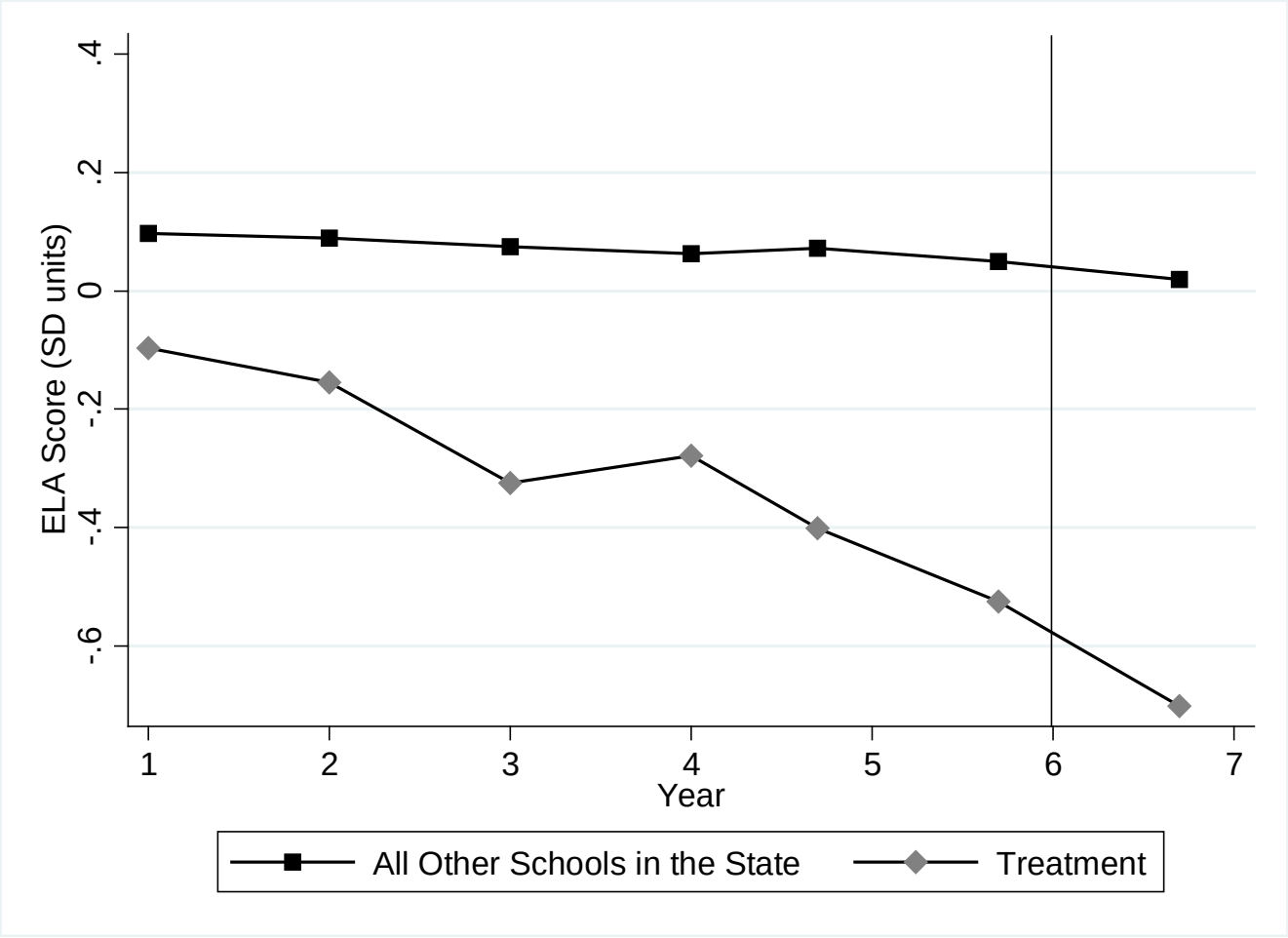
Math (All schools)



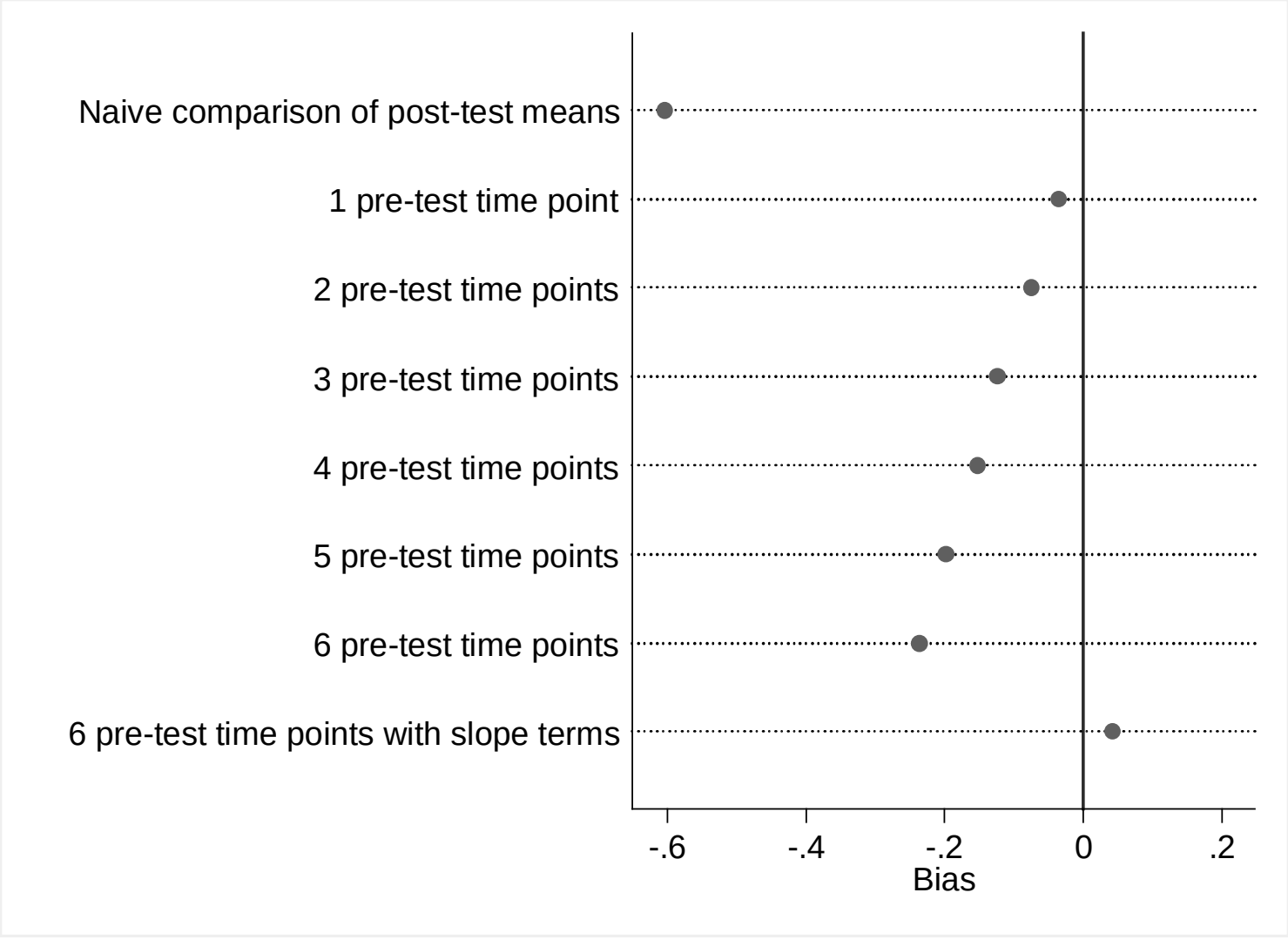
Math: WSC Results



ELA (All Schools)



ELA: WSC Results



With matching

- Same results
- Somers et al got the same results
- Environmental science found only replicate with matching

ITS Summary

- Comparative ITS does well relative to RCT to date
- Matching is the most consistent in getting results
- But models with the correct functional form do well and you can observe the functional form
- Need new studies in education – at least two are currently underway

MODELING A FULLY KNOWN SELECTION PROCESS

Statistical Theory

- Knowing the selection process and measuring it perfectly will always give you unbiased causal inference
- Rarely do we know it fully, but there are exceptions (Diaz & Handa, 2006)
- But often we have very strong information about the selection process even if it is not perfect – why children are retained in grade vs. why couples self-select into divorce

Known Selection Process: Diaz and Handa (2006)

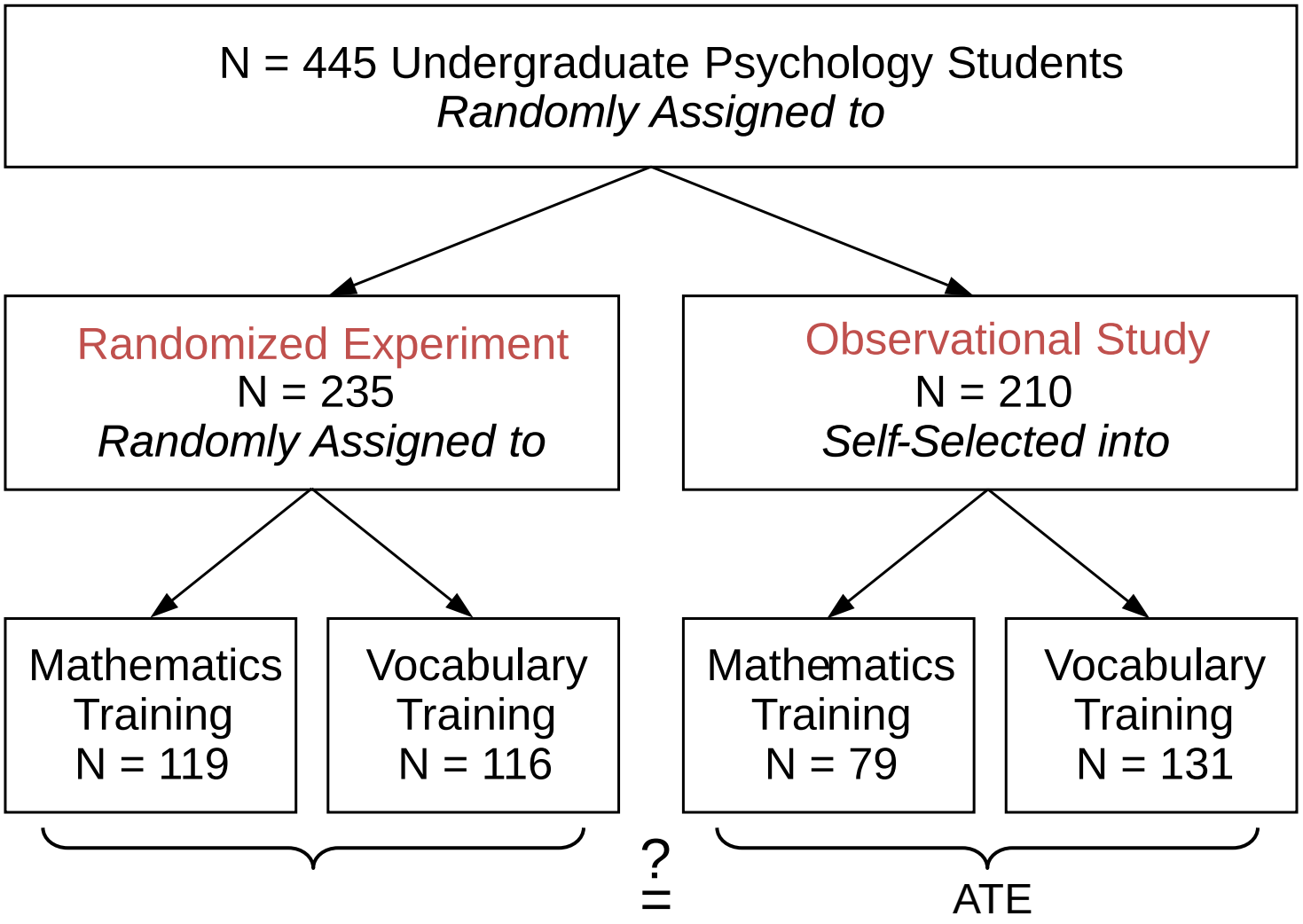
- RCT: Oportunidades
- Selection process: index of the material welfare of the home
- Quasi-experiment: going to more affluent villages, but where there are some residents that meet the selection standards for eligibility
- Analyze impoverished and less impoverished villages and found those with similar material resources get the same results as the RCT

Summary for Fully Known Selection

- Very rare to fully know the selection process
- Almost as rare to have no idea about the selection process
- Tasks are:
  - To gather as much information as possible about multiple variants of the selection process and measure them well
  - Or to select comparison groups up front that really minimize the unknown amount of selectiom

Possibly fully known 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)

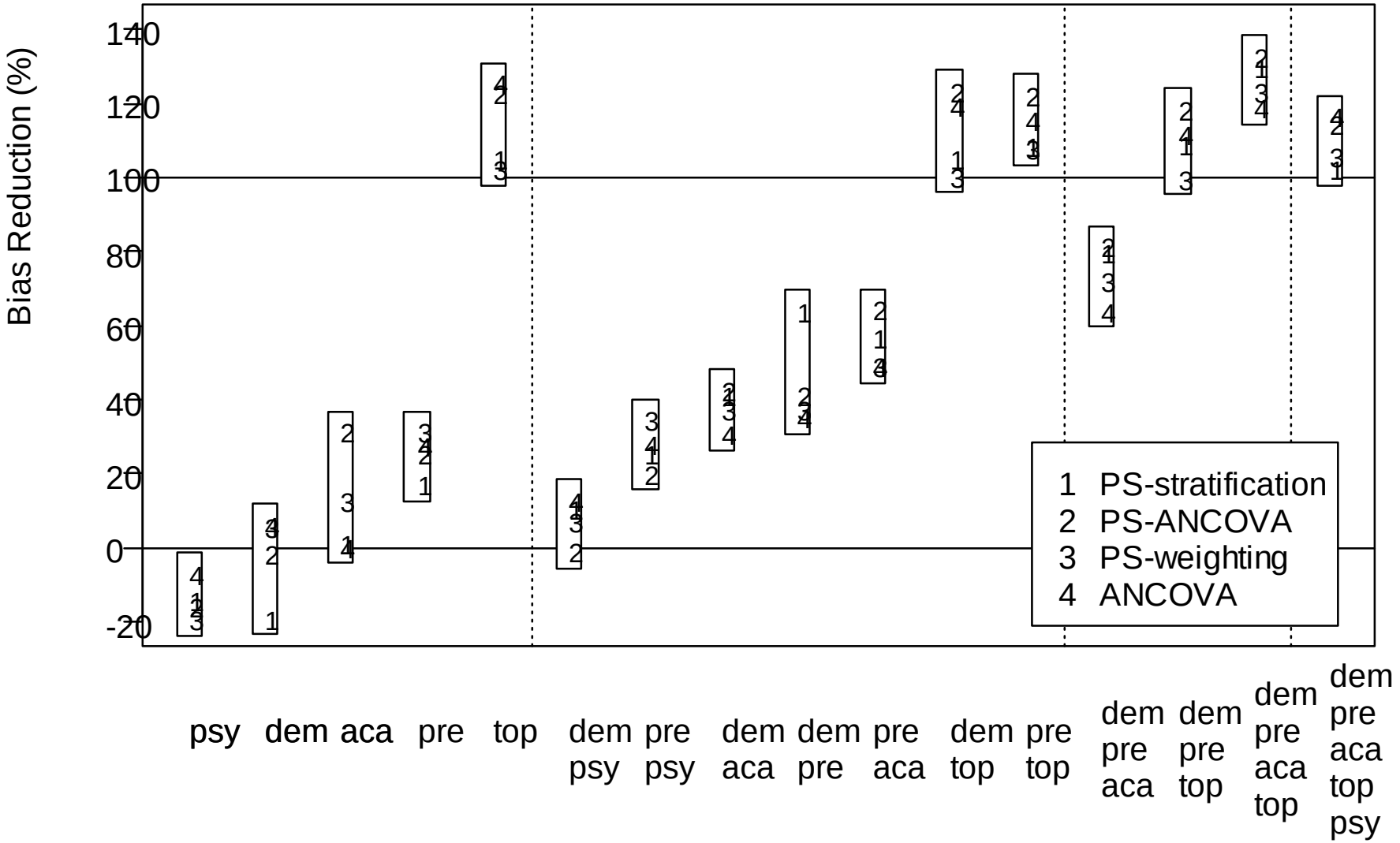
- *Demographics* (5 single-item constructs):

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

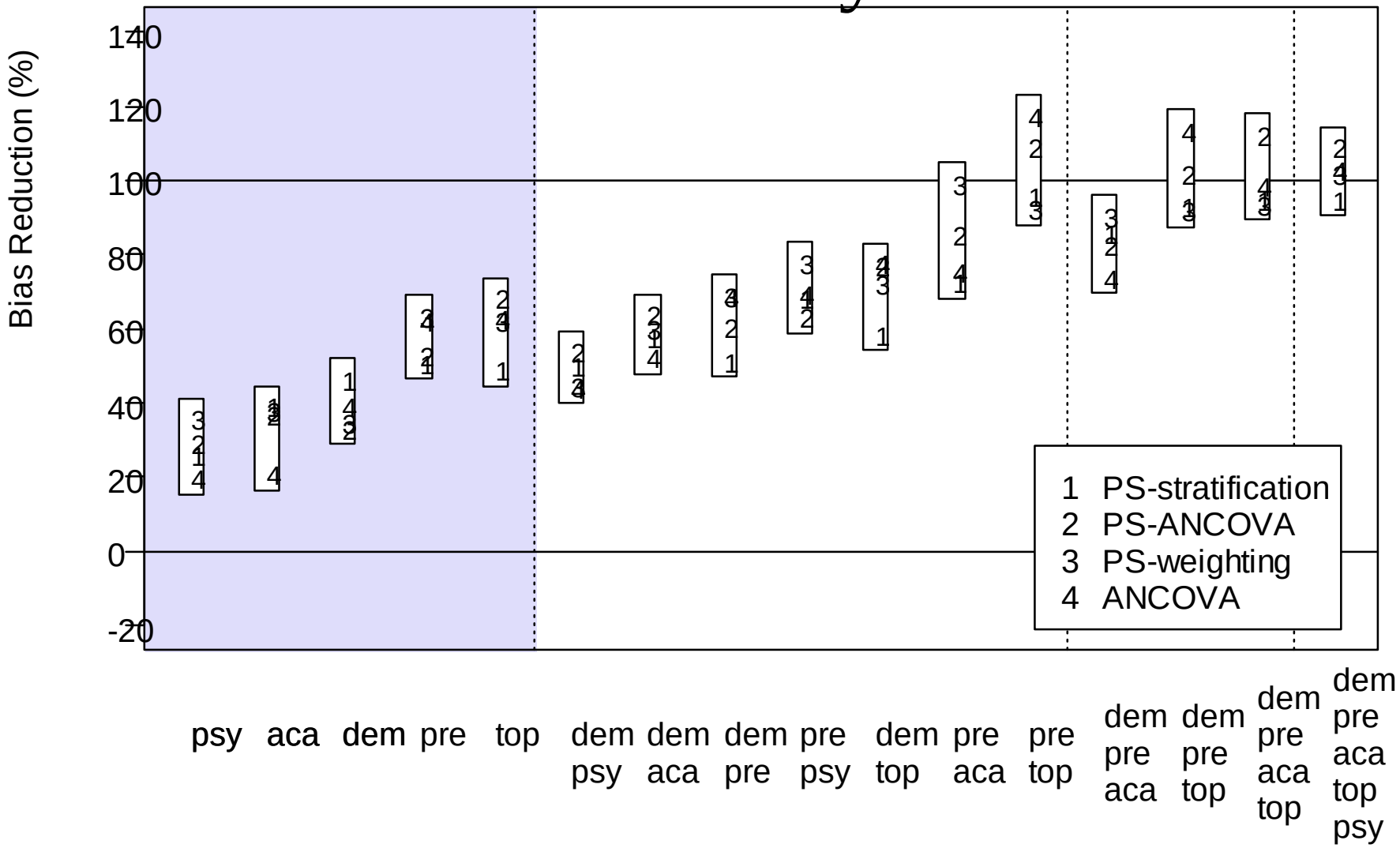
Was there Bias in the OS?

- Yes; math and vocab effects larger when students self-selected into respective T versus when there was random assignment
- Random assignment showed effects for each outcome.
- So our question is: How much of this bias is reduced by use of covariates? Or: can we recreate the estimates from the experiment?

Bias Reduction: Construct Domains Mathematics



Bias Reduction: Construct Domains Vocabulary



Question 2: What counts most for bias reduction given initial group differences?

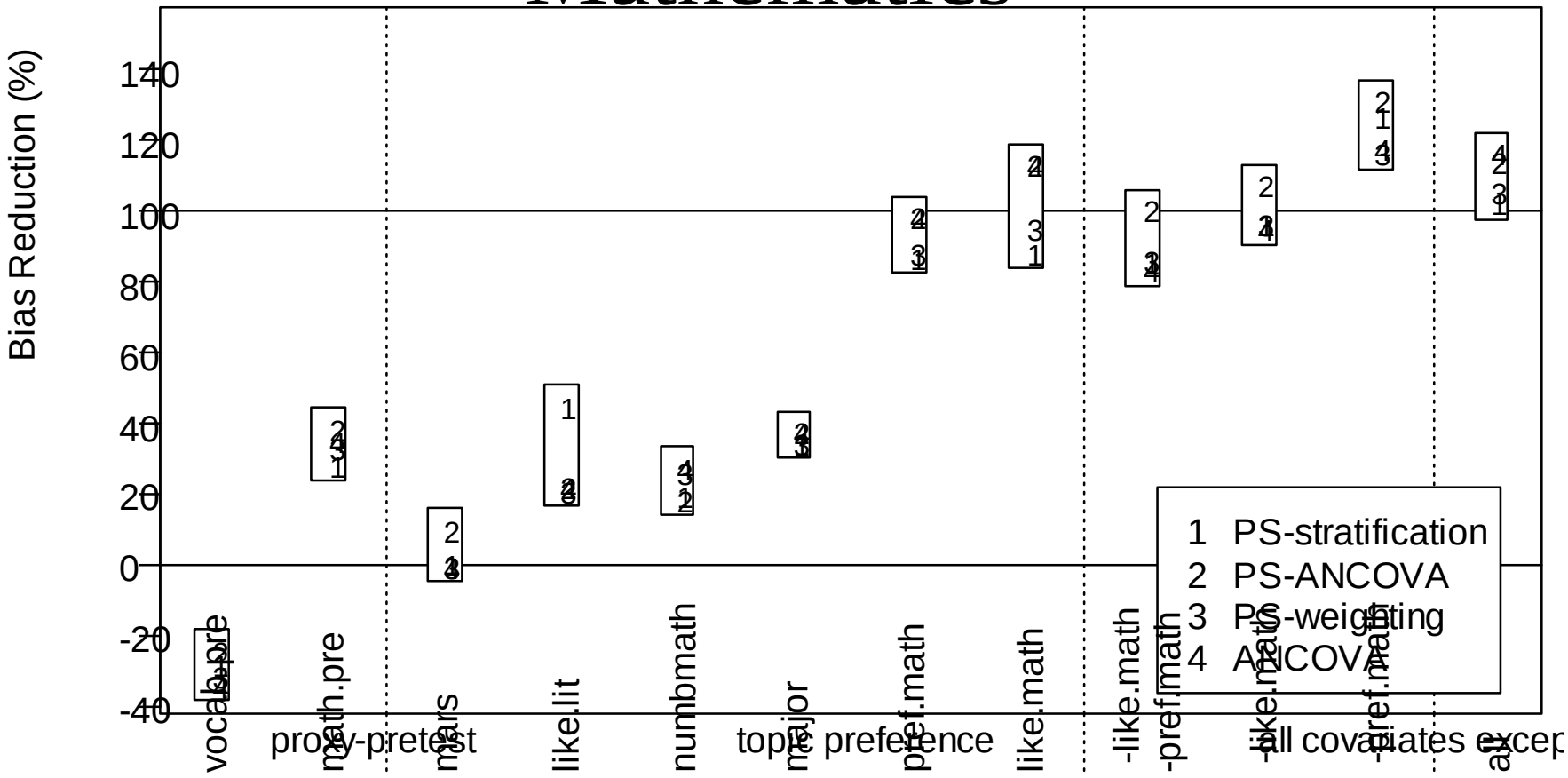
(a) Covariates measuring selection

(b) their reliability or

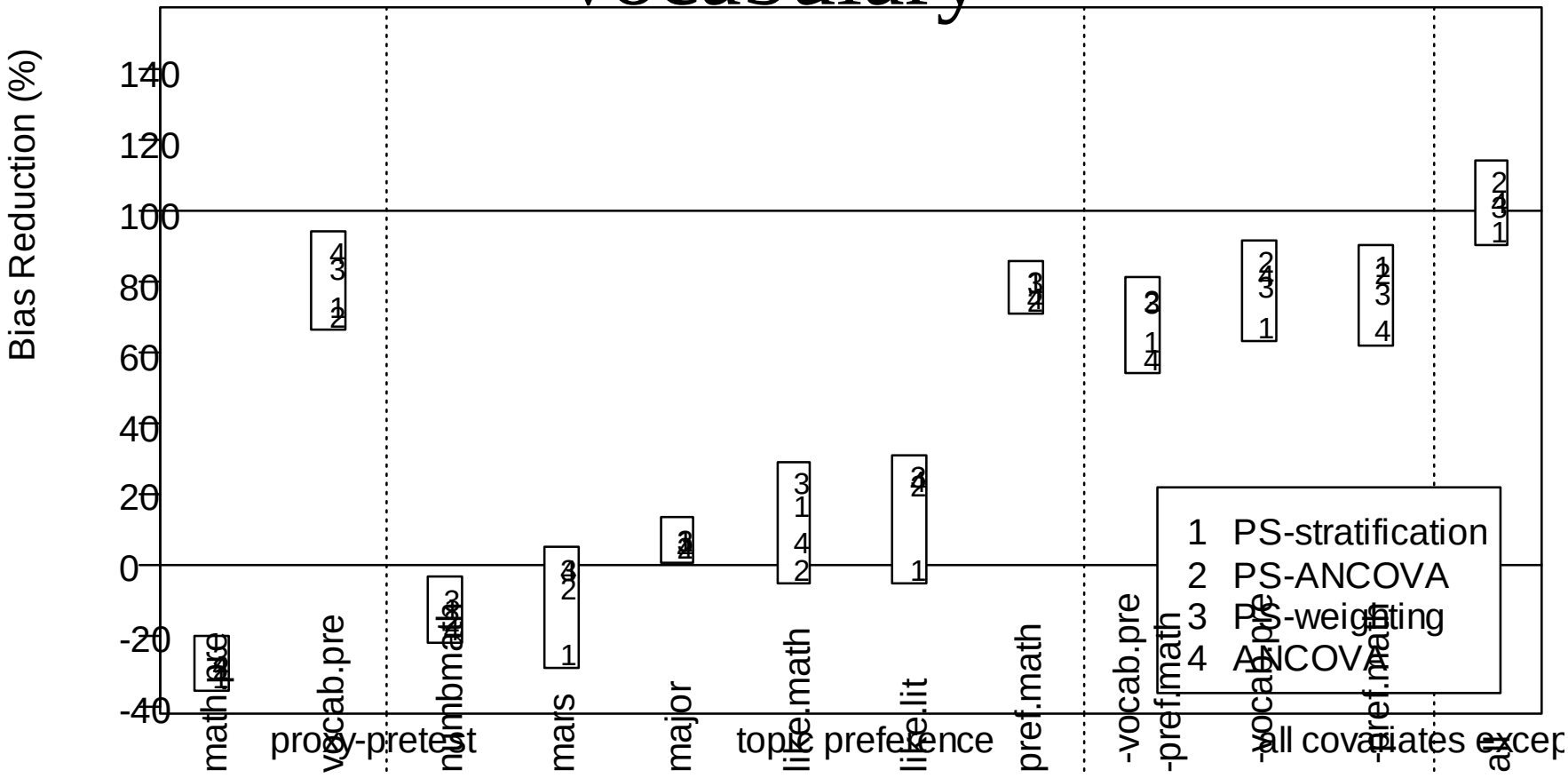
(c) how the OS data are analyzed?

(d) Use same data set, but has been replicated in Pohl et al. (2011)

Bias Reduction: Single Constructs Mathematics



Bias Reduction: Single Constructs Vocabulary

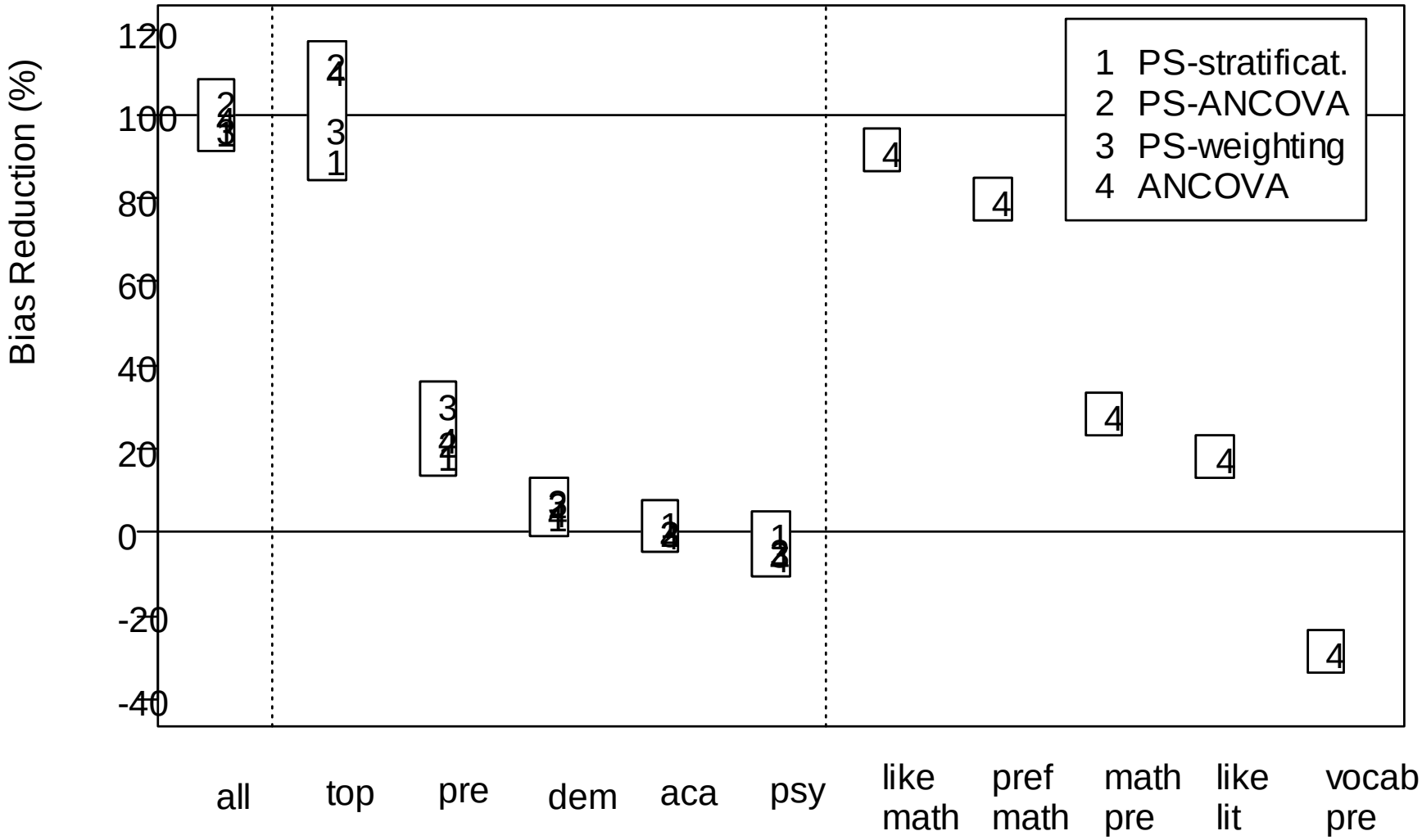


Reliability of Construct Measurement

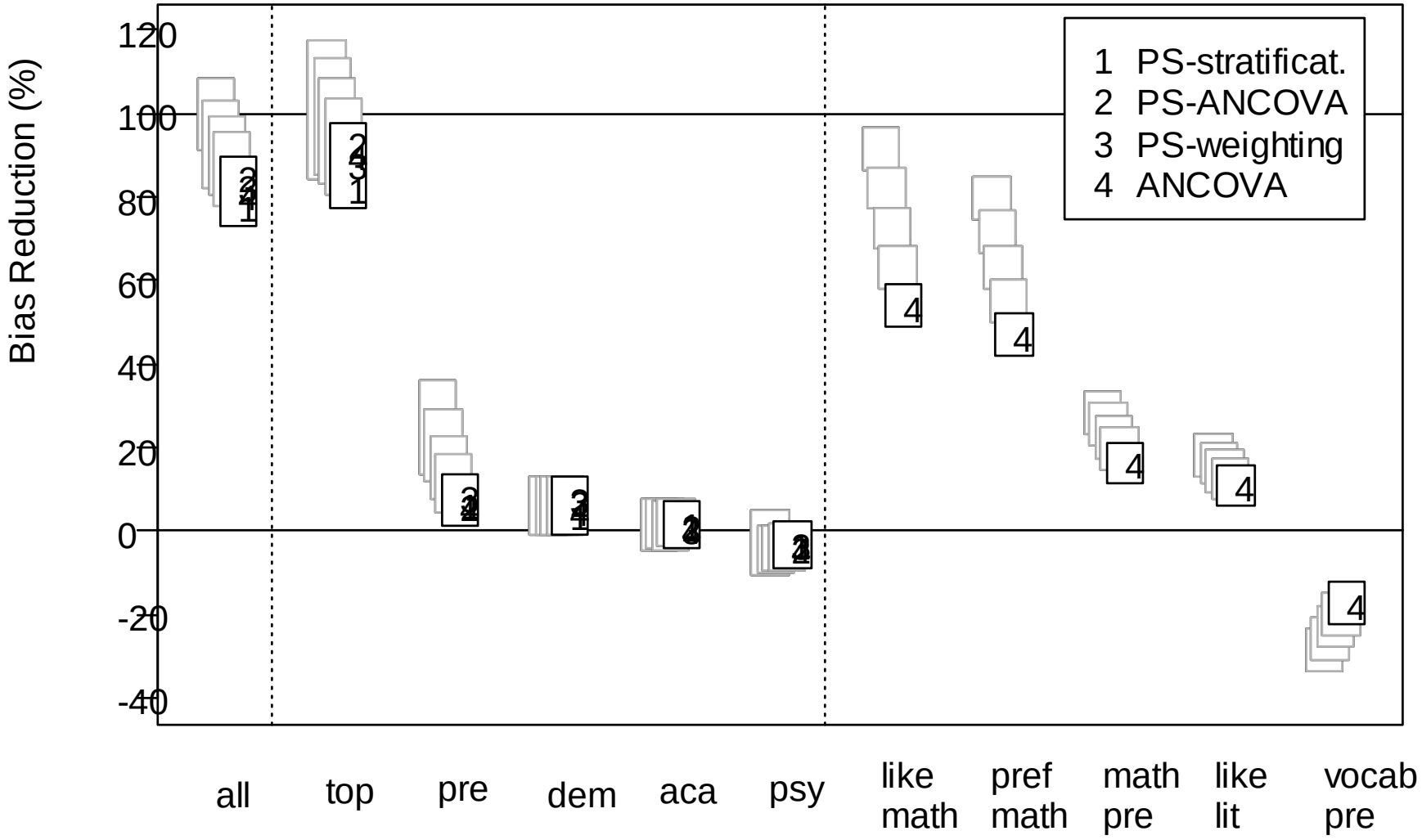
Steiner, Cook & Shadish (2011)

- How important is the *reliable measurement* of constructs (*given selection on latent constructs*)?
  - Does including many covariates in the PS model compensate for any one covariate's unreliability?
  - We add measurement error to the observed covariates in a *simulation study*
  - Assume that original set of covariates is measured without error and removes 100% of selection bias
  - Systematically added measurement error such that the reliability of each covariate was ρ = .6, .7, .8, .9, 1.0

Mathematics: Reliability 1.0



Mathematics: Reliability .6



Conclusions about Question 2: Relative priorities, given initial group differences

- The choice of covariates for selection adjustment is crucial
- Reliability counts, but is clearly secondary within bounds of 1 to .60.
- How you analyze the outcome using covariates (OLS and PS matching) makes little difference, though PS preferred in theory

PRETEST MEASURES OF OUTCOME: HOW SPECIAL IS THE PRETEST AMONG COVARIATES DESCRIBING THE WAYS NON-EQUIVALENT GROUPS DIFFER

Pretests and Proxy Pretests

- Campbell tradition has traditionally privileged designs with true pretest measures (Campbell, 1957; Campbell & Stanley, 1963; Shadish, Cook & Campbell, 2002)
- Warrant for special role
  - High correlation with the outcomes
  - Likely correlation with selection
  - Growth in longitudinal data systems has increased the availability

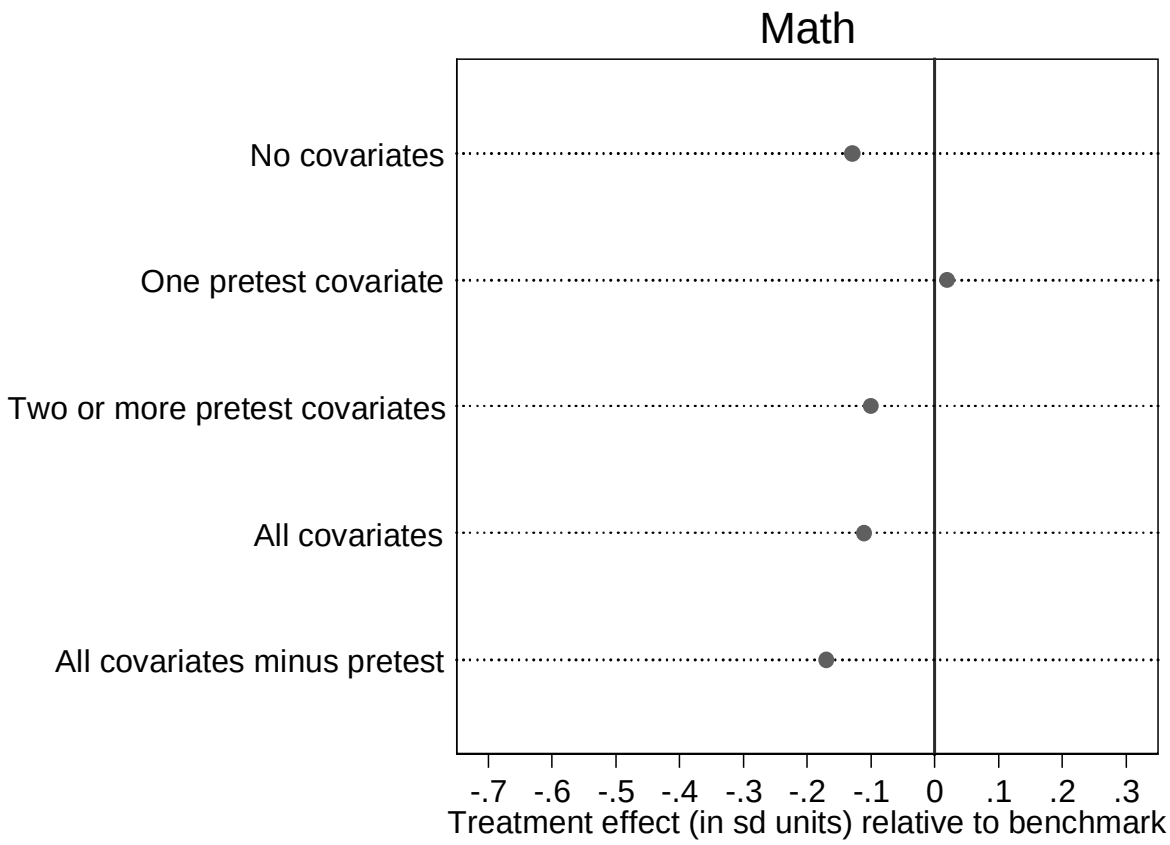
Existing Empirical Evidence

- Empirical WSCs provide some, but incomplete, support for privileging the true pretest in observational studies
  - Workforce development (Glazerman, Levy, & Myers, 2003; Bloom, Michalopus, and Hill, 2005; Smith & Todd, 2005)
  - Magnet school study (Bifulco, 2010)
- In both cases pretest is plausibly highly correlated with selection and outcomes
- This study examines the amount of bias reduction associated with conditioning on pretest measures when we vary the correlation with selection

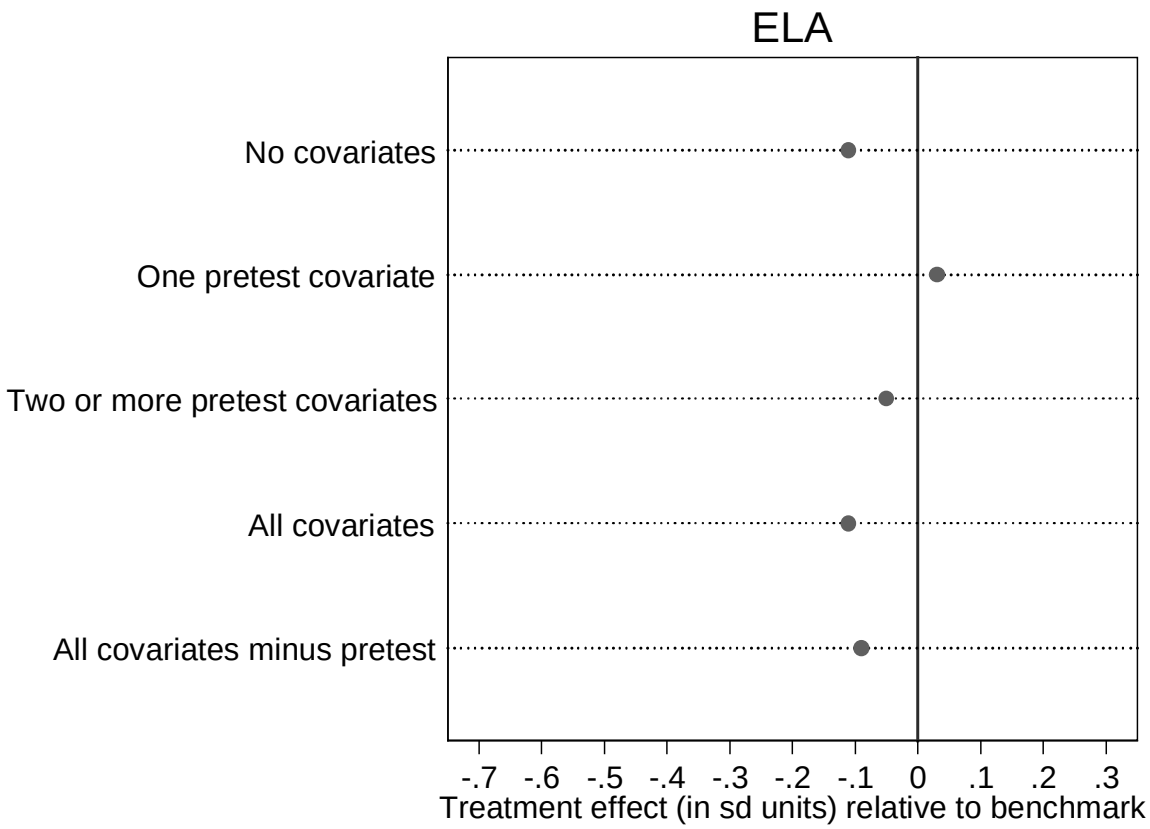
Case of little selection on the pretest

- RCT: Study of Indiana's system for feedback on student performance
  - Non-equivalent comparison group drawn from other schools in the state using a propensity score matching approach
- Idiosyncratic selection process
- Schools selected into the study because they were interested in implementing the program
- Principals interviewed and asked why they wanted to participate cited
  - Taking advantage of free resource from the state
  - A commitment to data driven decision making
  - Knowledge of other schools implementing
  - No explicit link between participation and the school's past academic performance

Bias Reduction in case of little selection on the pretest



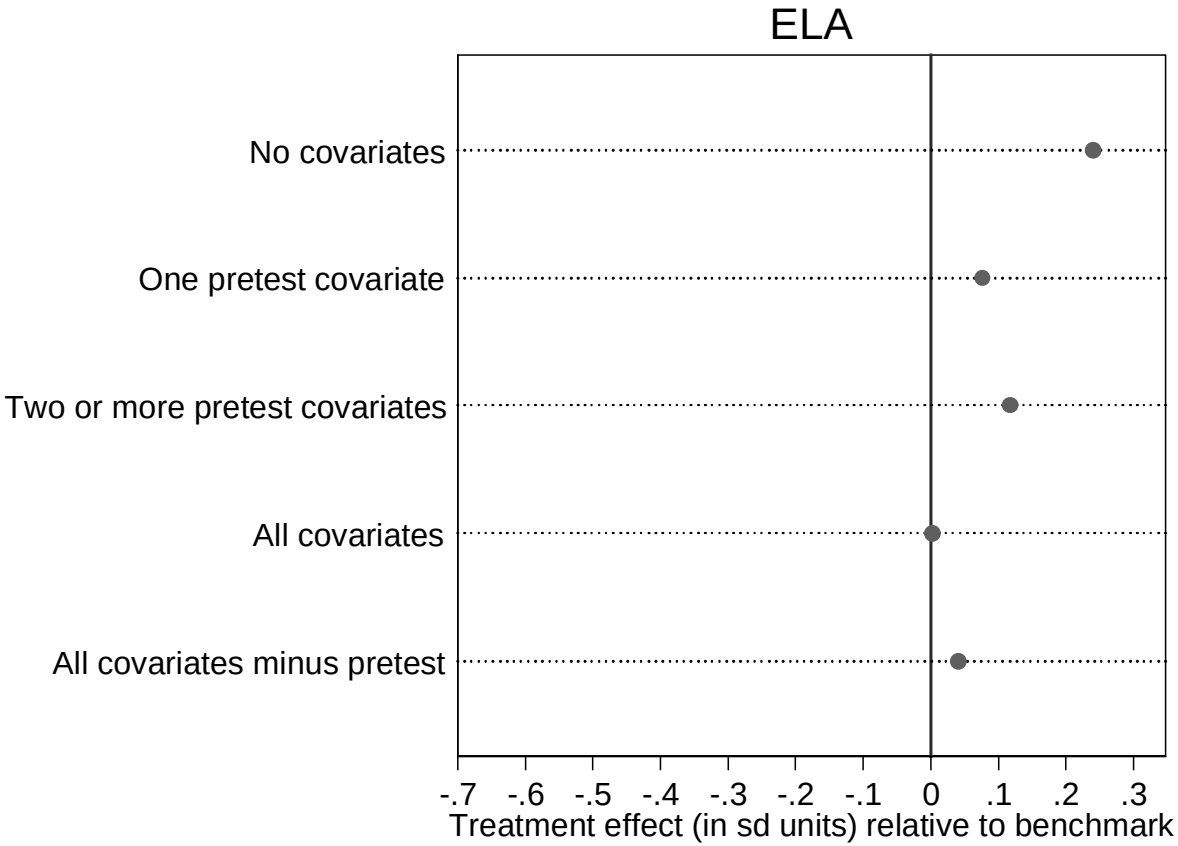
Bias Reduction in case of little selection on the pretest



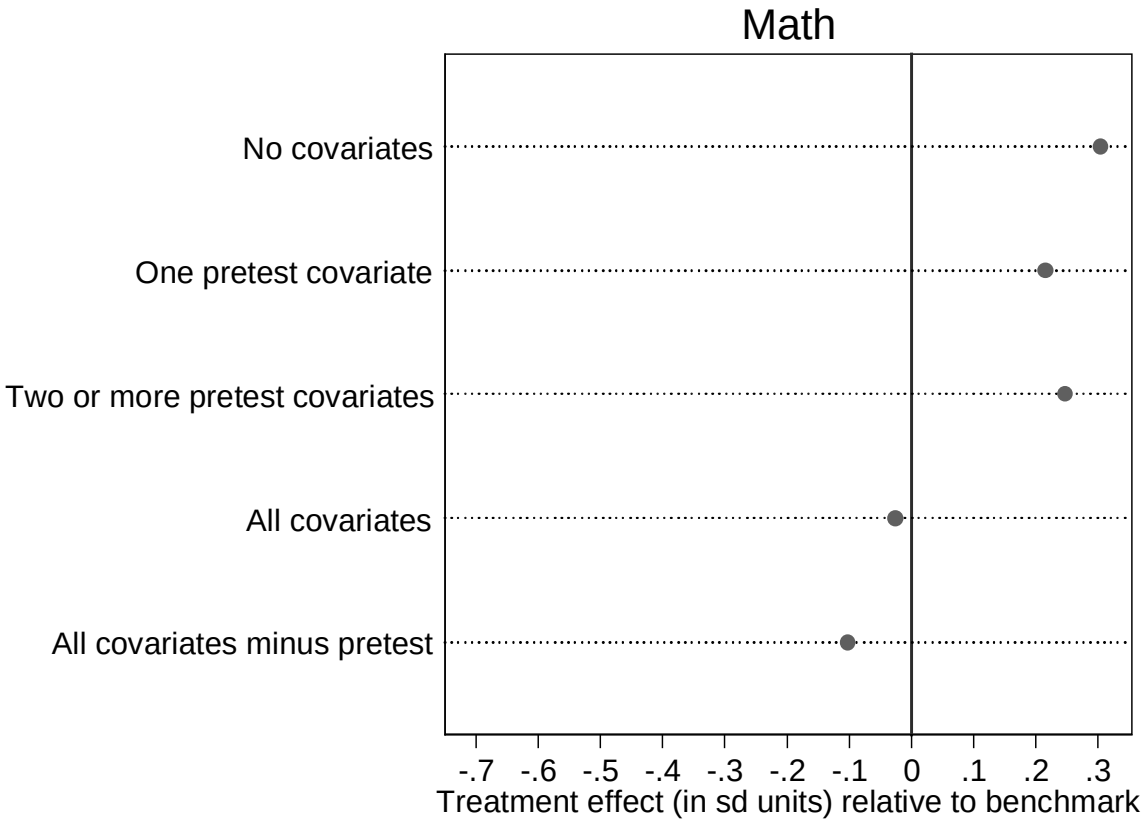
Selection on the pretest is differential

- RCT and QED: Shadish, Clark, and Steiner (2008)
- Reading pretest scores were predictive of whether students chose the vocabulary training
- This was not the case with mathematics

Bias reduction when selection on the pretest is differential



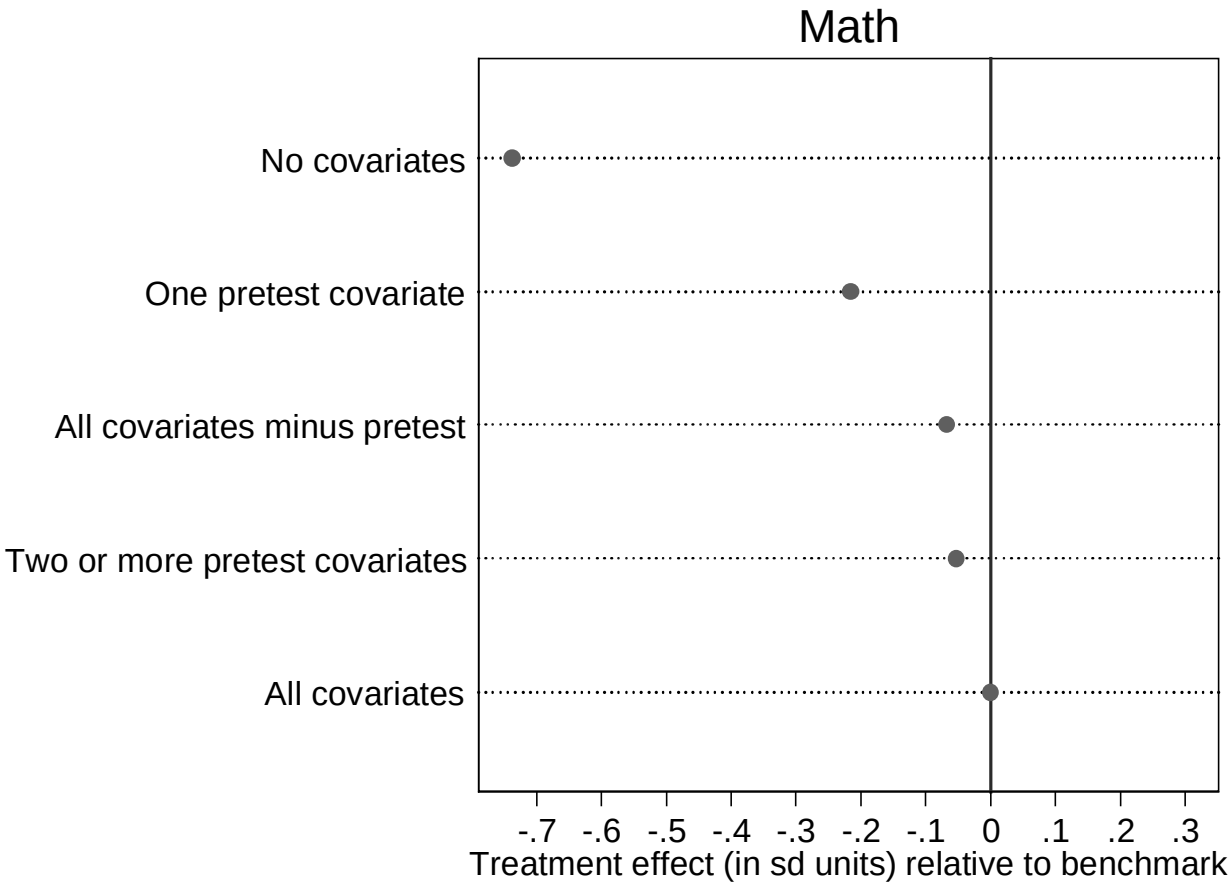
Bias reduction when selection on the pretest is differential



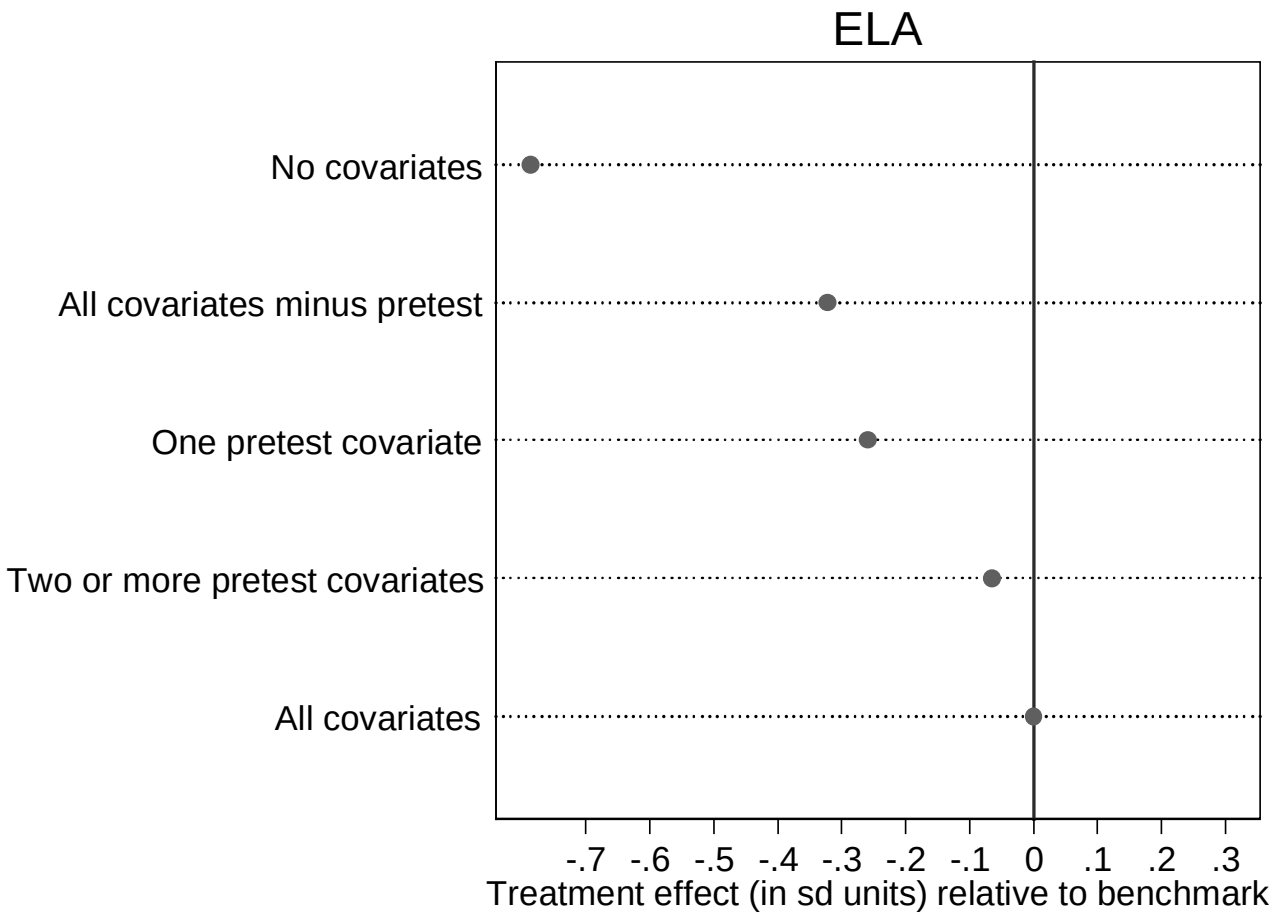
Selection on the pretest is clear

- Case study – kindergarten retention
  - Benchmark comes from Hong & Raudenbush (2005; 2006) who used the rich covariates in the ECLS-K to estimate the effect of kindergarten retention on academic outcomes in math and reading
- Past academic performance plays a critical role in identifying which students will be retained
  - Students are retained "*to remedy inadequate academic progress and to aid in the development of students who are judged to be emotionally immature*" (Jackson, 1975, p. 614)
  - "*It is a 'high risk' profile generally – for academic setbacks in the near-term, for a lifetime of struggle over the longer term.*" (Alexander, Entwisle, and Dauber, 2003, p. 68)

Bias reduction when selection on the pretest is clear



Bias reduction when selection on the pretest is clear



Summary of pretest results

- Cannot assume the pretest is always related to selection even though it may turn out that is very often related to selection
- Across the three dataset, we still find evidence that the pretest is an important covariate to consider in observational studies in education.
- However, the bias reduction associated with matching on the pretest is associated with how correlated the pretest is with *both* selection and outcomes
- If the pretest/selection correlation is weak, less bias reduction should be expected
- However, across all three datasets bias never increased as a result of matching on the pretest

Double pre-test design

- Could increase reliability
- Can get at time-varying selection processes if the measures are reliable
- If there a benefit from the double pretest, it will only be when measures are unreliable or the pretest functional form differences are very larges and generally captured by reliable
- More work is needed before final conclusions can be made about the importance of two pretest waves

SELECTING A NON-EQUIVALENT COMPARISON GROUP WHEN SELECTION IS NOT FULLY KNOWN: REDUCING ANY INITIAL NON-EQUIVALENCE

Guidance from the literatre

- Past empirical research suggests intact school matching performs best when:
  - Schools are matched on important predictors of selection and the outcome (focal matching)
  - Matches are geographically local (Cook, Shadish, & Wong, 2008)

Focal Intact School Matching

- Designed to select a set of comparison schools that are similar to the treatment schools on observable school-level characteristics known to be closely correlated with the outcome of interest
- Improves the plausibility of the strong ignorability condition
- Cook Shadish, & Wong (2008) identified three WSCs in which intact group matching led to similarity on pretreatment covariates of the outcome
  - In all cases, the QED estimates corresponded closely with the RCT benchmark

Local Intact School Matching

- Schools that are geographically proximal are often similar in both observed and *unobserved* ways
- Matching within district can be seen as analogous to including a district fixed effect in an OLS regression in that it rules out confounding from a correlation with the district-level error term and treatment
- Empirical WSCs in the job training literature and one education application have found local matches outperform non-local ones (Friedlander & Robins, 1995; Bell, Orr, Blomquist, Orr, & Cain, 1995; Heckman, Ichimura, & Todd, 1997; Bloom et al., 2005; Bifulco, 2012)

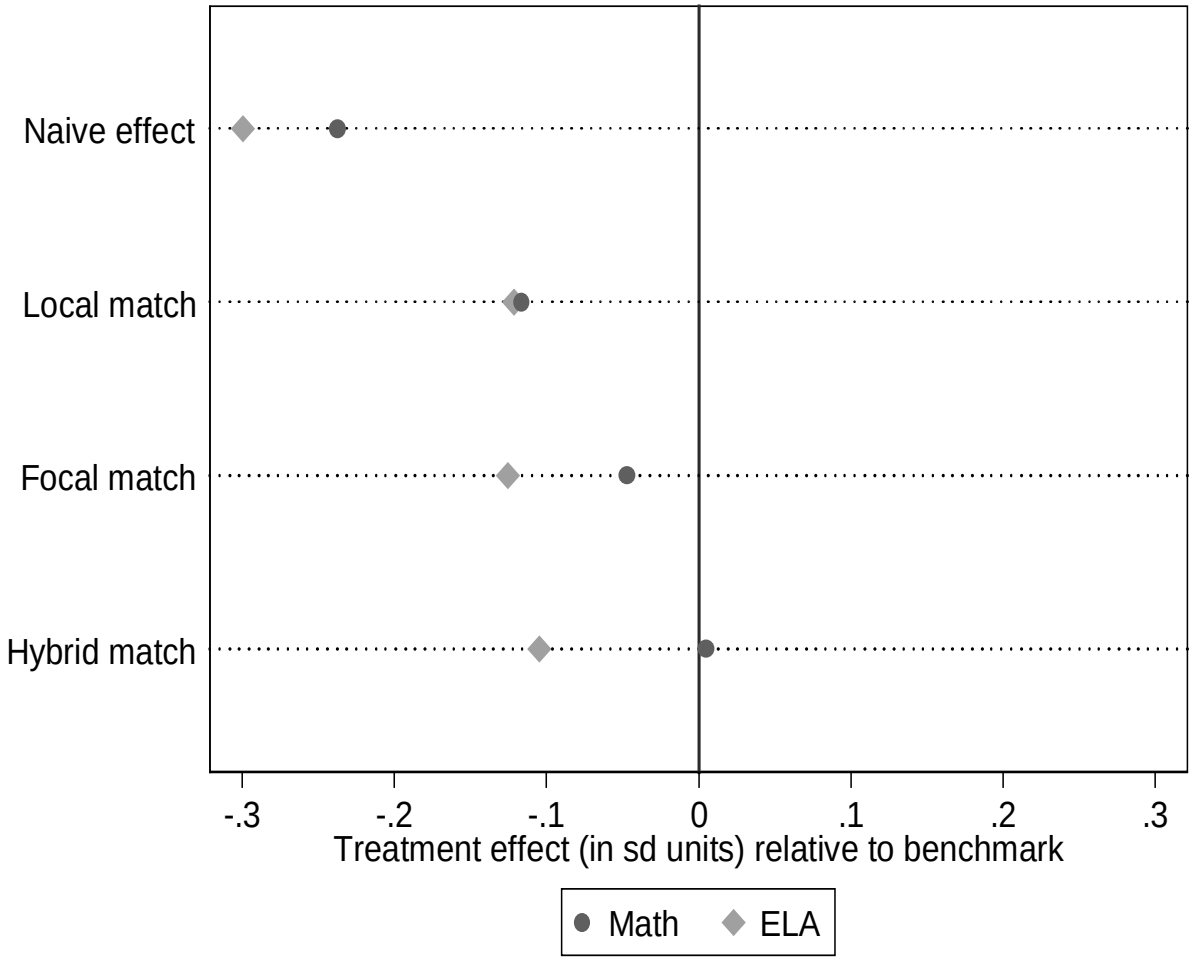
Potential Tradeoff between Local and Focal Matching

- In practice, finding a comparison school that is both geographically local and similar in important observable ways, may not be feasible (Stuart & Rubin, 2008)
- There is currently no guidance for researchers on the relative importance of finding a local or focal match
- Stuart and Rubin (2008) introduced a hybrid that preferences local matches unless they are outside of a pre-specified caliper on observable characteristics

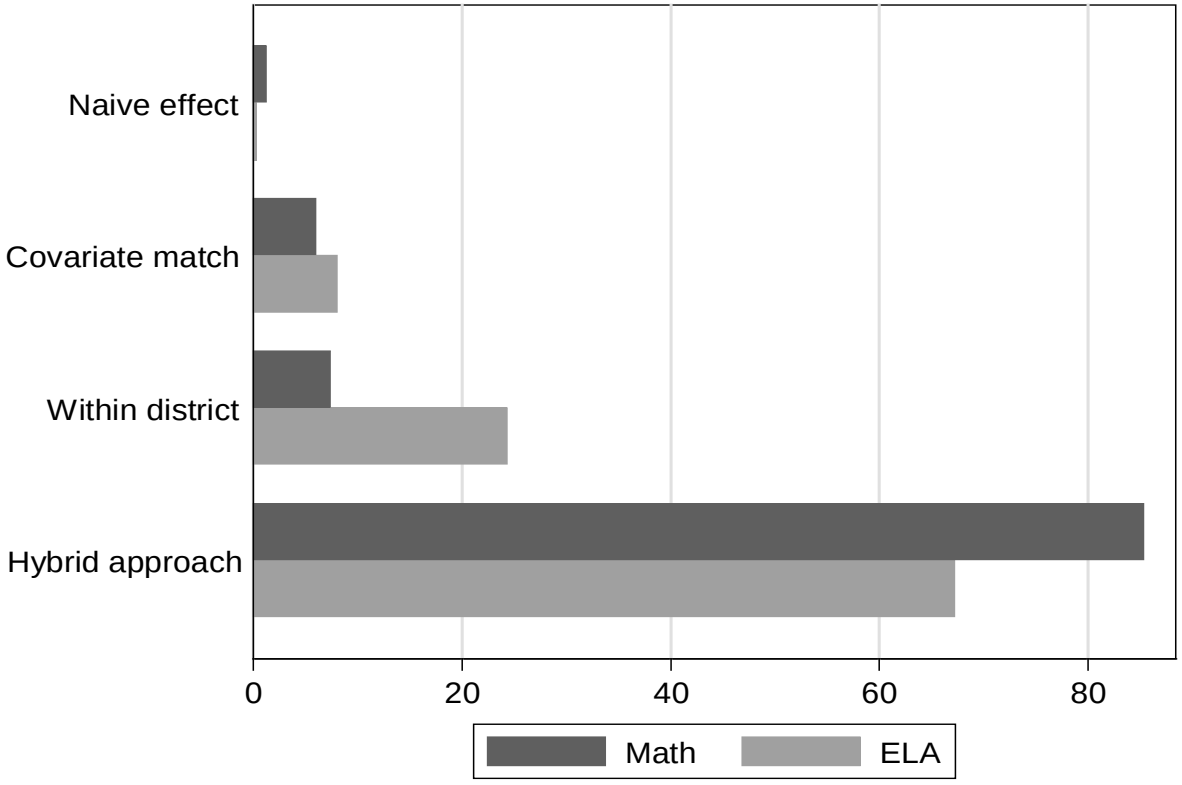
The test: Hallberg, Wong, & Cook

- This paper draws on a WSC to examines correspondence with the RCT benchmark (Indiana student feedback study) after matching
  - Within district (Local)
  - On observable school-level covariates known to be highly correlated with the outcome of interest (Focal)
  - Implementing a hybrid approach that preferences within district matches as long as the schools do no differ by more than 0.75 standard deviations of the propensity score (Hybrid)

Performance of naïve effect, local matching, focal matching and hybrid approach



Percentage of times observational approach performed best across 1000 replications



Summary

- Local matching and focal matching are generally good ideas, but they are not guarantees
- At present, it looks like Hybrid matching is best, but there is only one study, and it is at the school level rather than the individual level
- Need for more studies of hybrid matching, but there is reason to suspect it may be better than focal or local by itself

Overall Summary

- Although, there is not meta-analysis to date, things look good for RD, CRDD, and CITS
- It also looks very good when you design prospective studies and include measures to account for multiple possible selection processes
- 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 covariates

Overall Summary

- Focal and local matching each sometimes reduce all bias, almost always reduce some bias
- But it is likely they are best used together through a hybrid matching approach
- We need more studies for secure inference about the conditions under which non-equivalent control group designs reduce all bias
- Much work in this area is currently going on in different venues. 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.