Evidence-based Methods for improving Evidence-based Social Policy: Probing some Assumptions

Thomas D. Cook,
Northwestern University and Mathematica Policy Research, Inc.

Funded by NSF Grant DRL-1228866
The Within Study Comparison Method
WSC Design: Three-Arm Study

Overall Population sampled/selected into

Randomized Experiment randomly assigned to

Control Group  Treatment Group  Comparison Group

ITT of RCT  ITT of OS
WSC Design: Four-Arm Study

POPULATION
Randomly Assigned to

Randomized Experiment
Treatment
Control

Observational Study
Treatment
Control

ATE

? =
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 where an RCT is possible
- No reason to believe that a given design will always replicate experimental findings; our more modest goal is to identify designs that often replicate findings.
- This is inductive and requires a large sample size of WSCs. This talk is not the final word. More definitive word requires more WSCs
III. COMPARING RCT RESULTS TO REGRESSION DISCONTINUITY (RD) AND CRD
RD and WSCs

• 13 WSCs on RD at cutoff (Benchmark = RCT)
• 12 are 3-arm designs, and one 4-arm
• No M-A yet
• Impression is that the average LATEs are similar by design
• Impression is larger SEs
• These WSCs test the theory and, if M-A results are like the impressions reading the studies, then the theory of RD works in practice
CRD in WSCs

• Only 3 of these studies – we’ve briefly described before
• Test assumptions of parallel slopes with CRD-Pre and CRD-CG
• Find assumption holds in few examples to date if sample size large
• CRD clearly better than RD for effect and SEs
• CRD not clearly worse than RCT for both
WSC and CITS

• Four studies in medicine, three in education (one with 3 replicates) and one in environmental sciences
• All claim causal inferences similar
• 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 (schools as unit of assignment)
- Comparative ITS comparison groups
  - Basically all schools in the state
  - Matched schools in the state
Math: WSC Results

Naive comparison of post-test means
1 pre-test time point
2 pre-test time points
3 pre-test time points
4 pre-test time points
5 pre-test time points
6 pre-test time points
6 pre-test time points with slope terms
ELA: WSC Results

Naive comparison of post-test means
1 pre-test time point
2 pre-test time points
3 pre-test time points
4 pre-test time points
5 pre-test time points
6 pre-test time points
6 pre-test time points with slope terms

Bias
With matching C to T Units

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

• To date, CITS does well relative to RCT to date
• Matching is the most consistent to date
• But models with the correct functional form do well; one can observe the functional form
• Get similar effects despite possible group differences in (a) pre-treatment trend, (b) historical events at treatment; (c) changes in instrument; (d) stat regression— all these could be confounds, but they have not been to date
• CITS is not in any compendium except if matching for pretest equivalence
Alas...

• It is not common to know the selection process this well.
• However, it is rare not have any common sense knowledge of selection, at least
• Our general dilemma is not unknown selection process; it is partially known process
• What to do if you know you do not have needed variables from well known part of the analysis?
VI. MODELING A VERY WELL (BUT RARELY PERFECTLY) KNOWN SELECTION PROCESS
Diaz & Handa (2006): The RCT = select villages by poverty level, select households within village by scaled index of material hardship; randomly assign such households to T and C

- What is selection process into T?; is this deterministic?
- In villages that failed to get into RCT but were eligible.
- In villages that too rich to get into, but still had folks who were eligible.
Very well Known: I

- Shadish, Clark & Steiner (2008) – Will described processes to raise quality of knowledge about selection
- He was able to collect his own covariates in line with his 5 or so plausible selection models
- Almost all bias reduced
Very Well Known II: Hong & Raudenbush
WHEN SELECTION IS PARTIALLY KNOWN: FIRST, SELECT A NON-EQUIVALENT COMPARISON GROUP THAT MAXIMALLY REDUCES INITIAL NON-EQUIVALENCE BEFORE MATCHING INDIVIDUAL CASES
Intuitions about Comparison Group Choice

- The smaller the difference initially, the less the bias to adjust away
- The smaller the difference the greater likelihood that unobservables are controlled
- The smaller the difference the fewer cases will need to be eliminated (if match)
- So it is useful to select the comparison group carefully to increase overlap + generalization
Selecting an intact comparison group is sufficient to reduce even all bias

- (1) West et al (1999) – compare those signed up for RCT over summer with ACT 16 or below vs. the population of students not traced or deciding late for ASU, but also scoring 16 or lower and enrolled at ASU
- (2) Bloom & Michaelopolus – compare treatment sites in the same city with own control site versus all those in the control group in another site in same city
- Not balanced on unobservables; but are on those that count to affect the outcome
No guarantee selecting comparison group this way reduces all bias

- Consider at as the first stage in any bias reduction strategy
- Often overlooked in concern with matching cases. That lit. begins: Given a difference, what to do about it?; not minimize the difference up front, then deal with smaller
- Race to the Top in Kentucky – how to select
Importance of Local Comparisons

• Notice how sampling to increase overlap usually involves local comparisons, and how local may increase odds of dealing with hidden bias.

• In Shadish et al., non-equivalent comparison group was local – same university, age, course, time. Likely much overlap between T and C on correlates of selection and outcome, though differed in pretest scores and so further need for case matching.
Local Matching of Individual Cases
Focal Matching

• Focal matching: Product of analyses of variables responsible for selection and correlated with outcome – e.g., Shadish et al.– but depends on getting the right covariates
Hybrid sampling model of Stuart and Rubin (2008)

- Define a caliper for assuming an adequate match
- Take T and Match all LOCAL Cs within caliper
- For Ts without a C inside the caliper, perform a focal match with best propensity score one can predicated on analysis of selection processes
- A focal/local sampling hybrid results: (a) acceptably matched local Cs preferred for control over unobservables, and (b) acceptably matched non-local Cs, being matched only on observables
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

![Graph showing treatment effects for Math and ELA]
Percentage of times observational approach performed best across 1000 replications

- **Hybrid approach**
- **Within district**
- **Covariate match**
- **Naive effect**

Math and ELA categories are indicated by the bars. The horizontal axis represents the percentage of times each approach performed best.
Summary

- 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. Heckman claim.
- Neither is a guarantee.
- Hybrid matching is perhaps best, but only one study and at school and not individual level.
- Need for more studies of hybrid matching.
IX. WHEN SELECTION IS NOT WELL KNOWN, HOW SPECIAL IS A PRETEST MEASURE OF STUDY OUTCOME?
Claims about Pretest

• Claim that pretest is privileged for precision, but our concern is with bias reduction
• In bias-reduction strategies based on modeling the outcome, pretest often most highly correlated, but not always – job training
• For PSM, issue is correlation of pretest with selection into T
• Though selection on the pretest may be frequent, no one knows how often and when
• Next WSC studies vary when the pretest does and does not vary with selection
Existing Empirical Evidence

- WSCs provide strong support for privileging true pretest because it often reduces bias *(Workforce development)* (Glazerman, Levy, & Myers, 2003; Bloom, Michalopus, and Hill, 2005; Smith & Todd, 2005),

- and sometimes all the bias: Magnet school study *(Bifulco, 2010)* and earlier CITS studies here

- This study examines the bias reduction due to conditioning on pretest measures *when we vary the correlation with selection both between and within studies*
Between-Studies: Kindergarten Retention

• Hong and Raudenbush (2005; 2006) used the rich covariates in the ECLS-K to estimate the effect of kindergarten retention on academic outcomes in math and reading
Retention Selection Process

• 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)
## Dataset 1: Correlation with Selection

<table>
<thead>
<tr>
<th></th>
<th>Correlation with Retention in Kindergarten</th>
<th>Correlation Lower Bound</th>
<th>Percent of lower bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Pretest</td>
<td>-0.185*</td>
<td>-0.38</td>
<td>48.7%</td>
</tr>
<tr>
<td>Math Pretest</td>
<td>-0.179*</td>
<td>-0.37</td>
<td>48.4%</td>
</tr>
</tbody>
</table>
Data set 1: Analytic Approach

• Broke 144 possible covariates in three groups:
  – One wave of pretest data (measured spring kindergarten year)
  – Two waves of pretest data (measured fall and spring of kindergarten year)
  – All other covariates

• Created propensity scores with each set of covariates and estimated effects on reading and math, examined
  Bias reduction compared to benchmark model, not RCT!
Dataset 1: Math Results

![Graph showing treatment effects relative to benchmark with varying numbers of covariates.](image-url)
Dataset 1: ELA Results

- Treatment effect (in sd units) relative to benchmark
  - All covariates
  - Two or more pretest covariates
  - One pretest covariate
  - All covariates minus pretest
  - No covariates

ELA

Treatment effect (in sd units) relative to benchmark
Between- Study Contrast Dataset 2: Indiana Benchmark Assessment Study (Grade 5)

• 56 K-8 schools serving 5th graders randomly assigned to:
  – Treatment: implementation of the state’s benchmark assessment system (n=34)
  – Control schools: business as usual (n=22)
  – Outcomes: Math and ELA ISAT scores

• Quasi-experimental comparison group drawn from all other schools in the state that served 5th grade students (n = 681)

• Rich set of student and school covariates with multiple waves of pretest data
Dataset 2: Selection

• 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 – i.e., the pretest
Dataset 2: No Meaningful Correlation with Selection

<table>
<thead>
<tr>
<th></th>
<th>Correlation with Selection into Benchmark Assessment System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Pretest</td>
<td>0.041</td>
</tr>
<tr>
<td>Math Pretest</td>
<td>-0.012</td>
</tr>
</tbody>
</table>
Dataset 2: Math Results

Treatment effect (in sd units) relative to benchmark

- All covariates minus pretest
- All covariates
- Two or more pretest covariates
- One pretest covariate
- No covariates
Dataset 2: ELA Results

Treatment effect (in sd units) relative to benchmark

- No covariates
- One pretest covariate
- Two or more pretest covariates
- All covariates
- All covariates minus pretest
Why is there no Bias though Pretest is not correlated with outcome/
Shadish et al. Within-Study Difference in Selection

- Participants in QE arm asked why they selected the math or vocabulary training:

<table>
<thead>
<tr>
<th>Reason</th>
<th>Vocabulary</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liked the subject</td>
<td>18%</td>
<td>30%</td>
</tr>
<tr>
<td>Avoiding the other subject</td>
<td>21%</td>
<td>8%</td>
</tr>
<tr>
<td>Self-improvement</td>
<td>42%</td>
<td>47%</td>
</tr>
<tr>
<td>Good at the subject</td>
<td>17%</td>
<td>11%</td>
</tr>
</tbody>
</table>
## Correlation with Selection

<table>
<thead>
<tr>
<th></th>
<th>Correlation with Selection into Vocabulary Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Pretest</td>
<td>0.169*</td>
</tr>
<tr>
<td>Math Pretest</td>
<td>-0.090</td>
</tr>
</tbody>
</table>
Dataset 3: ELA Results where Pretest and Selection correlate

Treatment effect (in sd units) relative to benchmark:

- No covariates
- One pretest covariate
- Two or more pretest covariates
- All covariates
- All covariates minus pretest
Math Results where Pretest and Selection not correlate

- Treatment effect (in sd units) relative to benchmark
  - All covariates minus pretest
  - All covariates
  - Two or more pretest covariates
  - One pretest covariate
  - No covariates
Summary of Pretest Results

• Pretest is very often correlated with selection
• Cannot assume it is always related to selection
• When it is not correlated with selection, may be little selection to explain away and so little bias
• You should probably always include pretest, whatever Pearl says
• Heckman wants it included together with local comparison group matching and next subject
Little Explicit selection theory, but “Rich” set of covariates
Understandings of “Rich”

• Should be directly related to selection process and reliable
• Is usually numerous variables, but this silly
• We do not mean many measures of one construct not related to selection
• We do not mean unreliable measures of many constructs
• We mean reliable measures of a heterogeneous collection of constructs, given no strong theory
What is the “Theory” of multiple heterogeneous constructs?

• There is not much
• Gets at sources of hidden bias not captured by pretest or substantive theories of the selection process – Heckman
• Increases the chances of many modest independent correlations with the true selection process
• Not a good position to be in if no pretest, no theory of selection and measures derived from it
Worse Situations are:

• Having good understanding of selection process (even if imperfect), not having access to variables specified in model, but you go on anyway with whatever variables are on hand

• Having no idea of selection process and having few variables in one or a few homogeneous domains – e.g. demographics only or pretest only
Bias Reduction: Single Constructs
Mathematics

Bias Reduction (%)

proxy-pretest
vocab.pre
math.pre
mars
like.lit
topic preference
numbmath
major
pref.math
like.math
-all covariates except
like.math
-pref.math
like.math
-pref.math
all

1 PS-stratification
2 PS-ANCOVA
3 PS-weighting
4 ANCOVA
Steiner, Cook, Wei & Clark: Data

- Shadish et al. 157 covariates in 12 domains, examining 1, 3, 5 and 7 covariate items per domain. Outcome = learning
- Hong & Raudenbush. Effects of kindergarten retention on math and ELA outcomes; 208 covariates organized into 10 domains with 1,3,5,7 or 9 per domain
- In H&R, conclusions about bias not from RCT, but from an inferential argument: (a) retained students outperform retained, (b) retention at least partially due to test scores and teacher judgments of ability; and (c) using these as covariates reduces the diff between retained and promoted. So smaller difference = more bias reduction due to covariates.
“Effective” Covariates

• In SCS, analyses just presented: For math = liking for math and preference for math over vocabulary. Reported in Steiner & Cook

• In SCS, vocabulary pretest and preference for math over vocabulary

• In H&R, analyses in Hallberg (2013) that show almost identical causal estimates with 208 covariates and just pretests at 2 waves and teacher estimates of child ability at 2 times

• So take out both waves of these pretest values
Simulation Methodology

• 1. For any given size of domains, take all combinations to represent domains of, say, size 4 or 8.

• 2. For any given sample size of covariates within a domain, randomly sample from the pool of all covariates within that domain.

• 3. Do a 10 X 9 ANOVA for H&R and a 12 X 7 for SCS, in model testing for non-linearities.

• Present results all cov, and then minus best.
Results: All Covariates in Set

- Vocabulary (SCS)
- Reading (ECLS-K)
- Mathematics (SCS)
- Mathematics (ECLS-K)
Results minus Key Selection Covariates

Vocabulary (SCS)

Reading (ECLS-K)

Mathematics (SCS)

Mathematics (ECLS-K)
Summary of “Rich” Covariates

• It seems useful to conceive of “rich” covariates as domain heterogeneity and N of covariates within domains. May not be comprehensive, though.
• This is fallback; best is know selection process as well as possible, though never know it perfectly
• You really have to have better covariates in the set. H&R did modestly without, but SCS did not.
• Why did SCS not do so well here if it did earlier?
What about Multi-Level Matching?

• In Education, need to understand multi-level matching cos..

• Peter present his theory of it

• I will present results of one WSC on it
Slides to come