Within Study Comparison Workshop

Evanston, Aug 2012
What is a WSC?

• Attempt to ascertain whether a causal benchmark provided by an RCT is closely approximated by an adjusted QE.

• Attempt to ascertain conditions under which the approximation is closer.

• Attempt to make concrete what it means to have a perfectly known selection process.

• Attempt to provide empirically grounded fallback options for when an RCT is not possible.
Purposes of Workshop

• Learn more about the theory of WSCs
• Learn about some WSC implementations
• Increase rate at which WSCs are done now by coming up here with new issues
• Increase person power for doing WSCs by getting you motivated
• Challenge and extend the assumptions and practices about doing WSCs that our little theory group has adopted to date
Purposes of Today’s Discussion

• Review History of WSCs under 4 big rubrics: Empirical findings in different domains; Growth in types of questions asked; Growth in designs used; and Growth in criteria for WSC quality

  Suggest some issues still worth addressing
  Provide some input to begin process of deciding which work groups to be in
Quick Review of Empirical Literature
Pre-History of WSCs in Meta-Analysis

• Individual M-As compare average of RCTs and QEs done on a given topic
• Later M-As try to control for measured differences in design between the set of RCTs and set of QEs
• Meta-Analyses of M-As
• Basic Findings: 1. No or small mean differences by type of design
• 2. Greater variation in QE estimates than in RCT estimates in and across literatures synthesized
• What do we think of M-A as empirical method for testing design differences?
Earliest WSC History

• In domain of job training
• LaLonde. What he did and what he found
• Probably critical importance of a rapid replication – Fraker and Maynard
Subsequent History of WSCs in Job Training

- Meta-analysis by Glazerman, Levy & Myers
- Partial empirical summaries by Bloom
- and by Heckman group (Heckman, Smith and Todd)
- Factors that worked better (pretests and local comparison groups), though not well enough by the criteria adopted
The One Exception in Job Training: Bloom, Michelopolous et al.

Issue 1 in Bloom et al. is: Will the randomly and non-randomly formed control groups differ across 8 pretest observation points after standard statistical adjustments for any differences in mean, slope or other covariates.

• Issue 2 is: Rather than make statistical adjustments is it possible to sample non-equivalent comparison groups so as to eliminate, or at least minimize, the initial group differences one would otherwise have to adjust away.
The Context

- RCT is on job training at 11 sites
- Bloom et al restrict this particular analysis to 4 within-city comparisons where two RCTs in a city, thus matching on city characteristics, including job market ones, as well as on eligibility for job training and measurement of outcomes.
- Thus, these last factors are not confounded with treatment
- Here plot RCT control group and unadjusted QE group as short time-series – treatment a constant
Results: 3 within-city Samples
Portland--sample size smallest and least stable
Conclude?

• Seem to be some differences initially but not statistically detectable to Bloom et al.
• In 3 of 4 cities, sampling design alone minimizes initial differences. No difference in syntheses of all 4 cities.
• So in this application local controls matter. However,
• Cross-city comparison groups never gave similar RCT and QE results, despite 10 different analyses, incl OLS, propensity scores, Heckman selection models and random growth models.
• But number of pretest time points was related to the extent of bias reduction in cross-city tests, though never to a degree considered adequate.
Main Conclusion from Job Training
WSCs

- With WSC designs used to date, the statistical adjustments to the QEs have almost never recreated the RCT results.
- Why? Complexity of Selection Processes?
- Inadequacy of Covariates?
- Who knows and how to find out?
Under the Economists’ Radar at ASU

• In late 90’s, Aiken & West compare RCT, RDD and NECGD using a WSC
• RCT = remedial writing course at ASU, with pop = students within a narrow range on ACT, with outcome = multiple choice test and rated essays
• RCT result = reliable effect for knowledge of art of writing but not for actual essay quality
RDD is...

- Assignment variable is incoming ACT score
- Cut-off = score of 16
- Outcome = same as RCT
NECGD is...

• Some students applied late and were not around for assignment process
• Other students could not be contracted over summer when consents and everything for RCT were done
• NECG = all those from these 2 populations WHO ALSO SCORED BETWEEN THE ELIGIBILITY SCORES FOR THE RCT
Results

• RCT and RDD estimates very close, though not quite at same cut-off score, and reliable for multiple choice test but not rated essay
• RCT and NECG estimates very close and also differed across outcomes in same way as above.
• Few knew of this study; it contradicted main job training finding, should have implied caution about overgeneralizing job training WSC results. Some careful; others not.
Cook, Shadish & Wong: Reviewing WSCs with Comparable Findings

• RDD vs RCT at time: Aiken & West, Buddelmeyer & Skoufias; Black, Galdo & Smith
• RCT vs Sampling Matched Intact Groups: Bloom & Michaelopolous; Aiken & West; Diaz & Handa (1)
• RCT vs when selection process in NEGD is fully known (Diaz & Handa (2); Shadish, Clark and Steiner).
Diaz & Handa 1: Intact Matched Village Design

- Villages had eligibility criterion for RCT – bottom 20% on scale of material welfare.
- RCT compared eligible families within eligible villages that were or were not assigned to T
- Some eligible villages were not available for RCT for a variety of reasons, but were eligible
- Substituting eligible families in eligible but excluded villages same as eligible families in randomly selected villages
- So hidden variables played no role
Diaz & Handa 2: Non-Equivalent Village Sample

• Sample of non-equivalent and richer villages that included some eligible families
• Controlling for the set of material welfare covariates that fully determined family eligibility produced same results as RCT
• So any hidden bias due to differences in village affluence played no role
• Theoretically mundane to show no bias when selection fully explained (as in RCT and RD), but rich for advice about QE practice
Anomalies at the Time

- Agodini & M. Dynarski
- Wilde & Hollister
The Experiment is Project Star in 11 sites
- The non-equivalent comparison group formed from other Tenn. sites via propensity scores
- No pretest, but proxy background variables and some school data
- Analysis of 11 exp vs non-exp comparisons
- Conclusions are: (1) no equivalence in individual site comparisons of exp and non-exp results
- (2) pooled across sites, each significant but they differ in magnitude (.69 vs. 1.07)
What’s debatable here?

Design first: Who would design a quasi-experiment on this topic w/o pretest? with non-local and non-intact matches? Does this study compare a good experiment with a bad quasi-experiment?

Analysis: How good is a propensity score analysis with mostly demographic data?

• How valid is it to examine separate sites vs across sites?

• In any event, we have recovered the full cross-site treatment effect by first matching schools and then matching individual students within schools
• The experiment is on dropout at individual level at middle and high school
• The workhorse design uses propensity scores
• They are constructed separately from two sources—one another study at the middle school level and the other national data
• Findings: Few balanced matches are possible (29 of 128), given covariate data available and overlap achieved;
• Where balanced experiment and non-experiment do not produce same results
Commentary

How good was the experiment? 2 of 5 sites
• Control cases not from high school
• Testing at different times
• Pretest measures mostly not available
• How rich was the covariate structure? No theory of dropping out used, merely what was available in archive
• Modest exp contrasted with poor non-experiment
WSC Studies since then

• Many on RDD vs RCTs that Vivian will cover; strong conclusion is that they reach similar results
• Many on matching non-equivalent groups that Peter will review; some but not all reach similar conclusions
• A few that are emerging on RCTs and Interrupted Time Series (ITS) that Will will cover
The Shift in What Kinds of Questions WSC addresses
Are the Causal Estimates Comparable?

• In all areas of WSCs this was the dominant question until roughly around 2008
• There was some but not systematic interest in identifying the factors most responsible for getting comparable results with matching – e.g., role of the pretest and role of local matches.
• But this was relatively minor.
Shift in RDD Questions

• Most recent ones emphasize how the quality of implementation affects causal estimates or
• Probes how design supplements to the basic RDD help with functional form estimation, power and especially generalization beyond the cut point.
• These studies all have chance to replicate if estimates comparable at the cutoff under better conditions but that is no longer the main point
Shift in Matching Questions

• Attempt to identify the roles of covariate choice, reliability of covariate measures and mode of data analysis

• Attempts to identify which covariates are generally better: Pretests and the number and heterogeneity of covariates in general and when the better single covariates are omitted
No Shift in ITS Questions yet

• Newer area; and the ITS-specific tools for WSC work are still being ironed out
• The first data sets are just coming on board
• We have no idea if ITS and RCT estimates can be close
• But we will very likely get to a second wave of questions about the conditions under which estimates are closer and why some estimates are closer than others.
Improvements in WSC Design
3-ArmWSC Design

Overall Population
sampled/selected into

Randomized Experiment
randomly assigned to

Control Group
Treatment Group
Comparison Group

ITT of RCT \(=\) ITT of OS
4 Arm WSC: Shadish, Clark & Steiner

N = 445 Undergraduate Psychology Students
Randomly Assigned to

Randomized Experiment
N = 235
Randomly Assigned to

- Mathematics Training
  N = 119

- Vocabulary Training
  N = 116

Observational Study
N = 210
Self-Selected into

- Mathematics Training
  N = 79

- Vocabulary Training
  N = 131

ATE

? =

ATE
Advantages of 4 over 3 Arm Design

• Control over confounds from population differences, measurement differences, setting differences, etc.

• Compare this with LaLonde, pioneer but...
Limitations of 4 Arm Design

• Laboratory analog to what we usually want to generalize to; often lower external validity

• *May* limit the initial and subsequent group non-equivalence. Some commentators want to know how successful adjustments are as a function of the size of initial difference

• *May* limit the complexity of the selection process. Some will want to know how successful are adjustments as the selection process becomes harder to know
Provisional Advice is

• Try to replicate findings across internally more valid 4-arm design and externally more valid 3-arm designs after examining how good the latter are – more on the criteria for a better 3-arm WSC study soon.
Adding to Generalization

- Like all research, any one WSC finding may not be robust across different populations, settings, ways of assessing the treatment or of measuring the effect, and the times when the study is done.
- It will also have its own “limitations” with respect to the size of the initial non-equivalence, the complexity and current “knowability” of the selection process.
So we need...

• More WSCs done by researchers with different commitments
• More deliberate variations of the factors we think are responsible for closer or more distant approximations to RCT – exemplify via pretest data
• More empirical studies of selection processes in important research domains – again exemplify with pretest in H & R context.
Also, more Simulation Studies

• Steiner et al. adding unreliability to the data in Shadish et al.
• Open Discussion on Simulations
Improved Criteria for Evaluating WSC Studies
RCT is well Done

• Correct randomization procedure chosen
• Correctly implemented in field
• Pretest means not differ
• No subsequent attrition
• Remember Agodini and Dynarski
• Use covariates to control for sampling error in getting benchmark as point
• Guess this step is not done well in most WSCs
QE should be a good Example of its “Class”

• Comparing a good RCT to a poorly designed or implemented QE confounds how the comparison group is formed and quality of the RCT and QE

• Some stakeholders are more interested in how QE works when used with archival data. They argue that prospectively one can almost always use an RCT. So why test the prospective utility of optimally designed QEs when only the validity of archival QEs is of interest

• Their interest is not in QEs write large but in a sub-category of QE. They find the above requirement about QE quality to be of limited utility for the QE practice to which they aspire
Same Causal Quantity should be Estimated

- Don’t compare an ITT and TOT estimate, an ATE and a LATE
- Jeff Smith’s Paris critique in RD context
All Causal Confounds between RCT and QE should be controlled

• We are testing the effects of how the comparison group is formed, and nothing else
• Early studies confounded this with where comparison cases lived, when and how the outcome was assessed
• Now it is harder to get away with these things
• Unless one argues that one’s interest is in ability of QEs to control for both how groups are formed and where they live or when the outcome is assessed
Analysts of the RCT and QE should be blind to the Other’s Results

• Data Fishing to get the desired result far too easy
• Harder to do in practice than it sounds because RCT results are often well known
• Sometimes anomalous findings are due to silly mistakes by one analyst that need to be corrected
• Protocols should be developed in advance and compared to minimize problem above.
Criteria of Correspondence between RCT and QE should be clear

• Possibly the biggest single analytic problem, and we need more work since it is so central to WSCs
• Desire to treat mean RCT estimate as THE point benchmark, though it is subject to sampling error and so a causal region is more valid than a point
• More and more use of percent bias reduction relative to covariate-adjusted RCT
• Need to seek other options – like distribution match in multi-site RCTs??
Analysis of Generalization

• To publish, should we “insist on” at least a single replication though this is not a criterion in most other research – Fraker & Maynard?
• Is there a strategy for modified replications with sources of deliberate variation?
• Is there a role for simulations to extend knowledge of the conditions under which comparability of RCT and QE results is obtained or not?
Are these Criteria too Restrictive?

What are the chances of meeting all these quality WSC conditions in a 3 vs 4 arm test?

Is it desirable to seek to meet all these conditions? It limits QE practice to local controls, identical outcome measurement, etc. when some folks want to use distant controls and archival measures versus those directly collected in an RCT?
Next Steps:
Suggestions only
WSC Theory

• Do we need a paper examining and revising these criteria for a good WSC?
• Paper would struggle with, for example: (1) what are reasonable criteria for judging the correspondence of RCT and QE results; (2) how can blinding be promoted and adapted; (3) for which stakeholder groups are the current standards too restrictive? (4) other
WSC Design

• A critical examination of 3, 4 and 6 arm designs
• How to extend external validity in a practical way, given that extending it in WSCs is probably much like in other research
• Building simulations into WSC work.
• Is there a role for meta-analytic comparisons of RCTs and QEs?
• What is the role of directly observed selection processes, and how are they generalized?
• Other...
Tom’s little List

• Cohort designs minimize group non-equivalence, are prevalent and even more possible in education, yet we know little about how to deal with cohort differences.

• Selecting intact comparison groups before any attempts are made at case matching.
Regression-Discontinuity
Threats to Validity in RD

1. Correct specification of the functional form
2. Treatment misallocation
3. Potential confounds at the cutoff
Simultaneous designs

Aiken et al. (1998)
Synthetic RD Design
Gleason et al. (1012)
Four-arm design
Shadish et al. (2011)
Questions still Worth Asking About RE-RD WSCs

1. **Replication**: Does the RDD produce unbiased results, as compared to an experimental benchmark?

2. **Analytic**: Which analytic methods work in producing unbiased estimates in RD?
   1. Do current parametric methods work in modeling response function?
   2. Do current methods for bandwidth selection work in producing unbiased non-parametric estimates?

3. **Replication**: Does the RD replicate RE results in the field?

4. **Design elements**: Can design elements be used to detect threats to validity or strengthen designs?

   Can we use data from pretest, non-equivalent groups, or non-equivalent dependent variables to extrapolate beyond the cutoff?
Matching Questions: Peter
Observational Studies

• Issue with observational studies:
  – Deliberate selection into treatment conditions (self-selection, administrator selection, third-person selection)
  – Selection bias (confounding)

• If some stringent assumptions are met, unbiased treatment effects can be estimated
  – Strong ignorability
  – SUTVA
  – Methods & analytic assumptions (e.g. correct specification of models)

• Are these assumptions likely to be met in practice?
Three-Arm WSC Design

Overall Population sampled/selected into

Randomized Experiment randomly assigned to

Control Group  Treatment Group  Comparison Group

ITT of RCT  \(=\)  ITT of OS
Four-Arm WSC Design

Overall Population
*randomly assigned to*

Randomized Experiment
*randomly assigned to*
- Control Group
- Treatment Group

Observational Study
*deliberately selected into*
- Comparison Group
- Treatment Group

ATE of RCT

?= ATE of OS
Questions Still Worth Asking about Matching WSCs

1. *Replication*: can matching methods produce unbiased results, as compared to an experimental benchmark?
2. *Analytic method*: which analytic methods work best?
3. *Covariate selection*:
   – which covariates need to be measured?
   – which covariate selection strategies work best?
4. *Matching strategies* for multilevel data: which matching strategies work best for multilevel data?
5. *Design elements*: Can design elements detect threats to validity?
6. *Other purposes*
ITS questions: Will
RE vs Interrupted Time Series

- Series of observations on case interrupted by introduction of treatment (many variations)
  - Long vs short ITS
  - Rarely used in education except in some areas
- Least studied of all WSC’s
  - Only two examples I know, both in my lab
  - Anyone know of any others?
Randomized Longitudinal Crossover Design Approach: Roifman et al (1987)

• High-dose versus low-dose intravenous immunoglobulin in hypogammaglobulinaemia and chronic lung disease

• 12 patients in a longitudinal randomized cross-over design. After one baseline (no IgG) observation:
  – Group A: 6 receive high dose for 6 sessions, then low dose for 6 sessions.
  – Group B: 6 receive low dose for 6 sessions, then high dose for 6 sessions

• Outcome is serum IgG levels

• Here is a graph of results
Fig 1—Serum IgG concentrations.
A, initial high dose; B, initial low dose.
Meta-Analytic Approach: PRT

• Pivotal Response Training (PRT) for Childhood Autism
• This WSC method does a meta-analytic comparison of results from 14 studies using (66) SCDs to results from an RE.
• Meta-analytic WSC’s have a long history but also have significant flaws in that many unknown variables may be confounded with the designs.
  – But those flaws are often no more than in the usual 3-arm nonrandomized WSC
  – Big difference is the latter usually has raw data but meta-analysis does not. In the case of SCDs, however, we do have the raw data (digitizing data from graphs).
Initial Meta-Analysis (14 PRT SCD studies with ES aggregated to the study level)

------ Distribution Description ---------------------------------

<table>
<thead>
<tr>
<th>N</th>
<th>Min ES</th>
<th>Max ES</th>
<th>Wghtd SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.000</td>
<td>.256</td>
<td>3.564</td>
<td>.767</td>
</tr>
</tbody>
</table>

------ Fixed & Random Effects Model ---------------------------------

<table>
<thead>
<tr>
<th></th>
<th>Mean ES</th>
<th>-95%CI</th>
<th>+95%CI</th>
<th>SE</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td>1.0100</td>
<td>.8493</td>
<td>1.1706</td>
<td>.0820</td>
<td>12.3222</td>
<td>.0000</td>
</tr>
<tr>
<td>Random</td>
<td>1.4540</td>
<td>.9887</td>
<td>1.9193</td>
<td>.2374</td>
<td>6.1244</td>
<td>.0000</td>
</tr>
</tbody>
</table>

------ Random Effects Variance Component ---------------------------------

v = .592448

------ Homogeneity Analysis ---------------------------------

<table>
<thead>
<tr>
<th>Q</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>87.4789</td>
<td>13.0000</td>
<td>.0000</td>
</tr>
</tbody>
</table>

I² = 85.14%
RE on PRT: Nefdt et al. (2010)

• From one RE (Nefdt et al., 2010), we selected the outcomes most similar to those used in the SCDs
  – \( G = .875, \nu(G) = .146 \) (se = .382).

• Recall that the meta-analysis of PRT showed
  – \( G = 1.454, \nu(G) = .056 \) (se = .2374)

• Are they the same?
  – Same direction, somewhat different magnitudes
  – Again using 84% confidence interval overlap test:
    • \(.338 < .875 < 1.412\) for RE
    • \(1.120 < 1.454 < 1.788\) for SCDs
  • Again, the confidence intervals overlap substantially, so no significant difference between RE and SCD
RE vs ITS: Design Issues

• Is a four-arm randomized study feasible (attrition).
• Can we find data sets with RE and ITS on same question (Michalopoulos et al. 2004)
• Randomized Crossover Trials: Are there enough of these with many points over time.
• Meta-Analysis: Dealing with confounds of method with third variables (regression)
RE vs ITS: Analytic Issues

• What is the metric for the comparison of ITS and RE?
  – When the outcome is already the same over studies
  – When the outcome needs to be standardized
  – Both estimating the same parameter (ATE?)

• Dealing with trend in the ITS
  – Lots of issues and options

• Dealing with autocorrelation in the ITS
  – Bias in small time series
Questions Still Worth Asking about ITS vs RE

• The one area where we still need studies on the main effect question of “can ITS = RE?”

• Design variations to also examine:
  – Does it help to add a nonequivalent control?
  – Does it help to add a nonequivalent DV?
  – What about variations in ITS design?
    • Ordinary ITS with one intervention at one time
    • Multiple baseline designs with staggered implementation of intervention over time
      – Over cases
      – Over measures within one case
Work Groups

• Expectation is the thought-through outline of an IES proposal with question and its justification, data set(s) to be used, sampling plan, measurement plan, analytic approach, resources needed.

• How we meet Mon through Thursday – this can be modified. First time for us.

• Friday.