MATHEMATICA
Policy Research

Insert partner
or client logo
if needed

Considerations For Including Non-Experimental Evidence In Systematic Reviews

Mathematica Policy Research
Princeton, NJ

September 2016

John Deke

Considerations

- **Casual validity**
  - Is the observed impact real?
  - Primary focus of evidence review standards
  - Examples – attrition standard, baseline equivalence standard, RDD standards
- **Interpretation**
  - An impact on whom?
  - An impact of what?
- **Interpretable synthesis**
  - Synthesizing impacts in a way that is useful and accessible

Casual Validity
Is the impact real?

Conceptual Distinctions in Causal Validity

- **Key consideration for non-experimental designs – can bias be credibly bounded?**
- **Highest validity -- the estimated impact is due to one of two things:**
  - **Signal – a true effect of the treatment for the analytic sample**
  - **Noise – a random error that can be bounded with very high credibility**
- **Causal validity is threatened by systematic errors**
  - **Many systematic errors cannot be bounded – for example, self-selection bias**
  - **Some systematic errors can be bounded with pretty good credibility. For example:**
    - The WWC attrition bounds
    - Bounds that account for functional form misspecification bias in RDD studies that use an MSE-minimizing bandwidth

Randomized Controlled Trails

- The gold standard because it has the potential to achieve the highest level of causal validity
- This potential is often threatened in social policy experiments
  - Data are often missing for some of the randomized sample (attrition bias)
  - In cluster RCTs, the sub-cluster sample is often undefined at the point of randomization, creating the potential for unobserved sample selection bias
  - "Baseline" covariates used to improve precision in impact analyses are sometimes collected after randomization
- But these threats are somewhat manageable
  - Attrition standard, new standards on "joiners", and with late baselines at least it's attenuation bias

Matched Comparison Group Designs

- The original QED in the WWC
- Causal validity only achieved if an untestable and generally implausible assumption holds – equivalence of the treatment and comparison groups with respect to observed baseline variables guarantees equivalence with respect to unobservable variables too
- There is often little or no empirical support for this assumption; little or no theoretical support either

Three Original Categories of Causal Validity

Three Original Categories of Causal Validity

Low – QEDs and RCTs with high attrition that fail baseline equivalence standards

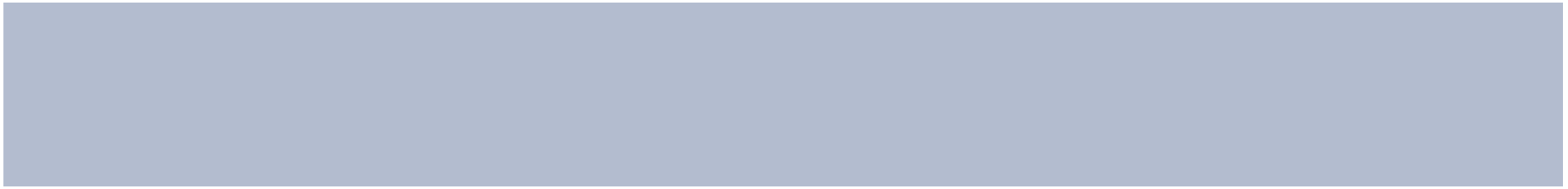
Middle – RCTs with unacceptable attrition that meet baseline equivalence standards; QEDs that meet baseline equivalence standards

High – RCT with acceptable attrition

Better Nonexperimental Designs

- There are nonexperimental designs that share some of the strengths of the randomized experiment
- Designs where assignment to treatment isn't random, but is very well understood (not a black box) and possibly even controlled by the researcher
- Examples
  - Regression discontinuity design (RDD)
  - Single case design (SCD)
  - Possibly others (interrupted time series, comparative RDD)

Ratings of Study Quality – Where Do the Best Non-experimental Designs Go?



Case Study: RDD

- RDD studies can attain the highest rating in the WWC – just like RCTs
- Argument against:
  - Though asymptotically unbiased, RDD impacts can be biased in small samples
  - Validity of RDD impacts depends on complex analyses
- Argument for:
  - RCTs with attrition or late joiners are also potentially biased
  - Techniques exist to bound small sample bias in RDDs
  - We were hopeful that standards could be developed to handle the complexity

Alternative Ratings Approaches for RDD

- **Alternative 1: Create a fourth rating category**
  - The fourth category could sit between matched comparison groups and pristine RCTs
  - Perhaps the attrition standard could be raised, and RCTs with moderate attrition could join RDDs and other non-experimental designs with plausibly boundable bias in this category
- **Alternative 2: Raise the bar for the middle category**
  - Make the highest category designs where errors in inference can be bounded with high credibility
  - Make the middle category designs where errors can be bounded with moderate credibility
  - Make the low category designs where plausible bounds are infeasible

Interpretation

An impact on whom?
An impact of what?

An Impact on Whom?

- Design/analysis methods can vary in the population/subpopulation to which an impact applies
- Is the impact an "average treatment effect" for the study sample?
  - The intent to treat (ITT) impact from an RCT
- Is the impact a "Local average treatment effect" for a subsample of the study sample?
  - The impact of "treatment on the treated" (TOT) in an RCT; also the "complier average causal effect" CACE
  - RDD impact at the cutoff of the assignment variable
  - Super local – fuzzy RDD impact at the cutoff of the assignment variable

An Impact of What?

- Exposure to the intervention (dosage), implementation fidelity, and the contrast with the control group can vary across studies
- It is less obvious that there are differences across study designs with respect to these issues

Interpretable Synthesis

Synthesizing impacts in a way that is useful and accessible

Challenges to Combining Impacts

- Impacts vary with respect to:
  - Causal validity
  - Population
- Implementation/contrast
  - These challenges are significant even with ITT impacts from RCTs
  - Adding more designs into the mix increases the challenge

Future

- **Standards focused on causal validity must continue to evolve**
  - Unlikely that standards will ever be "finalized"
  - State of the art methodologies continue to change, especially for nonexperimental methods
- **Challenges for evidence reviews**
  - Developing new standards for more non-experimental designs
  - Communicating to stakeholders the credibility of research findings from diverse designs
  - Communicating to stakeholders the meaning of research findings from diverse designs
  - Communicating to stakeholders that research evidence is not written in stone