Can Quasi-Experiments Yield Causal Inferences?

Matthew L. Maciejewski, PhD
Durham VA HSR&D and Duke University

Sample

	Year	Age	Race	SES	Health status
Study 1					
Study 2					

Intervention

	Interventionist	# of interactions	Duration of each interaction	Single topic or multiple topics	Content
Study 1					
Study 2					

RCT considered Gold Standard of Benefit Design for Several Reasons

- Create balance in observed covariates
  - Reduces number of competing hypotheses for variation in outcomes to one (treatment assignment)
  - Control group outcome is a valid counterfactual (unbiased estimate of outcome for treatment group had they not been randomized to treatment)
- Treatment effect generalizes to entire sample
- Statistical result is causal effect of treatment on outcome

Context for Perceived Inferiority of Quasi-Experiments

- Prior comparisons of RCTs and non-RCTs
  - Experimental results rarely replicated
  - Even when applying instrumental variables (IV) methods (LaLonde 1986)
- RCTs typically compared to non-identical samples and non-identical outcomes in different data
  - Conclusion has been that design (quasi-experiment) is the cause of difference, not sample or outcomes
  - Could outcomes be similar across designs if same sample & outcomes?

Differences in Samples for RCTs and Quasi-Experiments

- RCTs
  - Conducted on highly selected populations
  - Rarely pregnant women, highest risk people, oldest
- Quasi-experiments
  - Conducted on general populations
- Differences not necessarily due to randomization
  - Could be entirely due to different samples included

LaLonde (1986) Job Training Results

Estimator	Wage Difference for Men
Unadjusted RCT	$886
Non-RCT estimates from PSID & CPS-SSA	
Unadjusted	Low=-$1637, High=$1714
Age adjusted	Low=-$1388, High=$195
Age, schooling, race & pre-period wage	Low=-$1228, High=$1466
IV	Low=-$667, High=$889

Stukel 2007 JAMA: Mortality Impact of Cardiac Catheterization

Model	*Risk Ratio (95% CI)*
Unadjusted survival	0.36 (0.36, 0.37)
Multivariate adjustment	0.51 (0.50, 0.52)
Simple PS Adjustment: Deciles + Covariates	0.52 (0.51, 0.53)
Fancy PS Adjustment: Deciles + Covariates	0.52 (0.51, 0.53)

Conclusion

1) Adjustment for covariates important in non-RCT
2) Multivariate & PS regressions are same

Stukel 2007 JAMA: Mortality Impact of Cardiac Catheterization

Model	*Risk Ratio (95% CI)*
Unadjusted survival	0.36 (0.36, 0.37)
Multivariate adjustment	0.51 (0.50, 0.52)
Simple PS Adjustment: Deciles + Covariates	0.52 (0.51, 0.53)
Fancy PS Adjustment: Deciles + Covariates	0.52 (0.51, 0.53)

What to conclude?

1) Regression & PS results are both right?
2) Results are both wrong?

Stukel 2007 JAMA: Mortality Impact of Cardiac Catheterization

Model	*Risk Ratio (95% CI)*
Unadjusted survival	0.36 (0.36, 0.37)
Multivariate adjustment	0.51 (0.50, 0.52)
Simple PS Adjustment: Deciles + Covariates	0.52 (0.51, 0.53)
Fancy PS Adjustment: Deciles + Covariates	0.52 (0.51, 0.53)
Instrumental Variables	**0.84 (0.79, 0.90)**
RCT Results	**0.79-0.92**

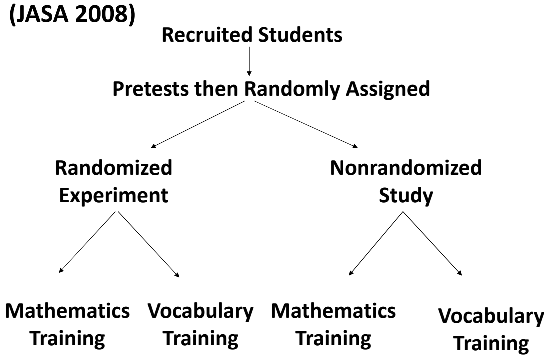
What to conclude?

1) Regression & PS results are both right?
2) Results are both wrong? **This is it.**

Re-appraising the Value of Quasi-experiments

- An under-used design allows direct comparison of results from RCT & non-RCT
  - Within-study comparison study
- Four-arm study: 2-stage process
  - Randomize to randomized treatment or self-selected treatment
  - Same treatments, controls, outcomes, timing
- Can compare two treatment effects!
  - Difference btn treatment & control in RCT "arm"
  - Difference btn treatment & control in non-RCT "arm"

Design of Within-Study Comparison by Shadish (JASA 2008)



Details of Shadish (2008) Design

- Participants from one college
- Participants pretested on several covariates
- Chose math and vocabulary training because
  - Easy to induce effect with item difficulty
  - Math phobias cause plausible selection bias
- All participants treated together (in same class) without knowledge of different conditions
  - People randomized to math in same training class as people self-selecting math
- Everyone post-test on math & vocab outcomes

Unadjusted Results:

Vocabulary Training Effect on Vocabulary Outcome

	Vocab Training	Math Training	Mean Difference	Absolute Bias
Unadjusted RCT	16.19	8.08	8.11	
Unadjusted Quasi-experiment	16.75	7.75	9.00	0.89

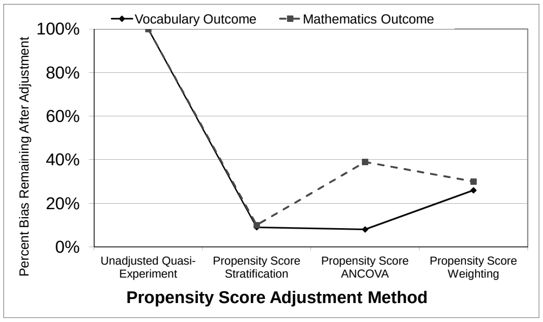
Conclusions

1. Effect of vocab training on vocab scores was larger (9 of 30 points) when participants could self-select into vocabulary training
2. The 8.11 point effect in RCT was overestimated by 11% (0.89 points) in the quasi-experiment

Propensity Score Modeling

- Based on a priori model of selection process that informed prospective pre-test assessments
- Extensive adjustment
  - Math & vocabulary pretest scores, ACT, GPA, prior exposure to math courses, math anxiety, demographic
  - "Big 5" personality traits (extraversion, emotional stability, agreeableness, intellect, & conscientiousness)
  - Extensive adjustment reduced bias a lot (59-96%)
- Limited adjustment (comparable to claims)
  - Age, sex, race & marital status had reduced bias modestly (12-30%)

Bias Reduction Fairly Similar Across Different Propensity Score Methods



Implications of Shadish (2008)

- Sampling design produced non-equivalent groups on observables
- Big overlap in baseline values in RCT & non-RCT groups due to 1st stage randomization made propensity scores more valid
- Extensive measurement of relatively simple selection process, though not homogeneous
  - Propensity score matching may not be effective if selection process is complex (as in job training)
- Bottom line: Propensity score results from extensive adjustment matched RCT results

Limitations of Shadish (2008)

- Short duration (15 minutes)
  - Not costly to conduct
  - Little incentive for non-compliance
- Absence of non-compliance with treatment assignment
- Short time between pretest & post-test, and short time between treatment & posttest
  - Change attributable to few things besides treatment
- Not generalizable to complex medical settings
  - Longer duration, have significant non-compliance and delay between treatment and outcomes assessment

Conditions Under Which Quasi-Experiments Match RCT Results

- Similarity between groups in pre-period values
  - When geographically local, comparison groups may not differ on major observables b/c provider & site effects controlled (e.g., pts in same clinic)
  - ACEI example (Hebert & Maciejewski)
- Rigorous conceptualization and measurement of selection process to support effective matching
  - Pre-period outcomes are particularly important
  - Adjustment using "off the shelf" vars not enough
- Regression discontinuity

Descriptive Statistics of Unmatched CA ACEI and Non-CA ACEI Cohorts

	CA ACEI Cohort?	Non-CA ACEI Cohort?	Standardized Differences
Age	76.1	75.9	7.17
Female (%)	65%	64%	2.09
White Race (%)	76%	83%	17.41
Black Race (%)	9%	7%	7.38
Baseline AMI (%)	6.7%	4.1%	11.52
Elixhauser Score	5.69 (7.79)	4.66 (6.99)	44.54
Baseline Expenditures	$8081(15210)	$6180 (12798)	16.06
Baseline # Meds	6.7 (3.8)	6.4 (3.6)	26.03
Office visits	8.6 (8.3)	8.5 (9.0)	4.42

Careful Consideration of Selection Process

- Bias can be significantly reduced if three steps of confounder adjustment are done
  - Identification of all relevant confounders from literature, theory, and experts
  - Error-free measurement
  - Proper modeling
- Use of variables of convenience fails 1st step, so unlikely to reduce bias fully
  - Especially true in claims data?

Reconsider Value of Quasi-Experiments for Causal Inference?

- Comparing good RCT to poor quasi-experiment confounds design type and the quality of its implementation
  - Logical fallacy
- This conclusion is ex post facto because we know RCT results in advance
  - Rarely true; more often have to infer ala Stukel
- Quasi-experiments satisfying three conditions more likely to generate valid causal estimates

Questions?

References

- Cook, Shadish & Wong, 2008. *J Policy Analysis and Management*, 27(4): 724-750
- Diaz & Handa, 2006. *J Human Resources*, 41(2): 319-345.
- LaLonde RJ, 1986. *Amer Ec Rev*, 76(4): 604-18
- Shadish, Clark & Steiner, 2008. *JASA*, 103(484): 1334-1356
- http://steinhardt.nyu.edu/scmsAdmin/uploads/002/477/Tom%20Cook-FINAL.pdf

References & Resources

1. Cook, T. (Producer). (2008). When Experiments and Observational Studies give comparable Causal Estimates:. [Powerpoint presentation] Retrieved from http://steinhardt.nyu.edu/scmsAdmin/uploads/002/477/Tom%20Cook-FINAL.pdf
2. Diaz, J. J., & Handa, S. (2006). An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator. [Article]. Journal of Human Resources, 41(2), 319-345.
3. LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. The American Economic Review, 76(4), 604-620.
4. Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. Journal of the American Statistical Association, 103(484), 1334-1344. doi: doi:10.1198/016214508000000733
5. Stukel, T. A., Fisher, E. S., Wennberg, D. E., Alter, D. A., Gottlieb, D. J., & Vermeulen, M. J. (2007). Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. JAMA, 297(3), 278-285. doi: 10.1001/jama.297.3.278