Non-Experimental Approaches for Evaluating Educational Interventions

John A. Gonzalez, Ph.D.
Rackham Graduate School
University of Michigan
Michigan Association of Institutional Research
November 5, 2015

How do we know if a program works?

- Compared to what?
- What are the levels of “working” vs “not working”?
- For whom?
- Does “working” require meeting goals?
- What should be considered acceptable evidence for making evaluative decisions? Causal Claims? Expert Opinions? Stakeholder Consensus?

Why Are Experiments the Gold Standard?

- They eliminate alternate explanations
- They help establish temporal order (I see a change in Y by manipulating X)
- They allow us to isolate effect of treatments without confounding baseline characteristics (theoretically)
- They help us tease-out effects by focusing on a single point

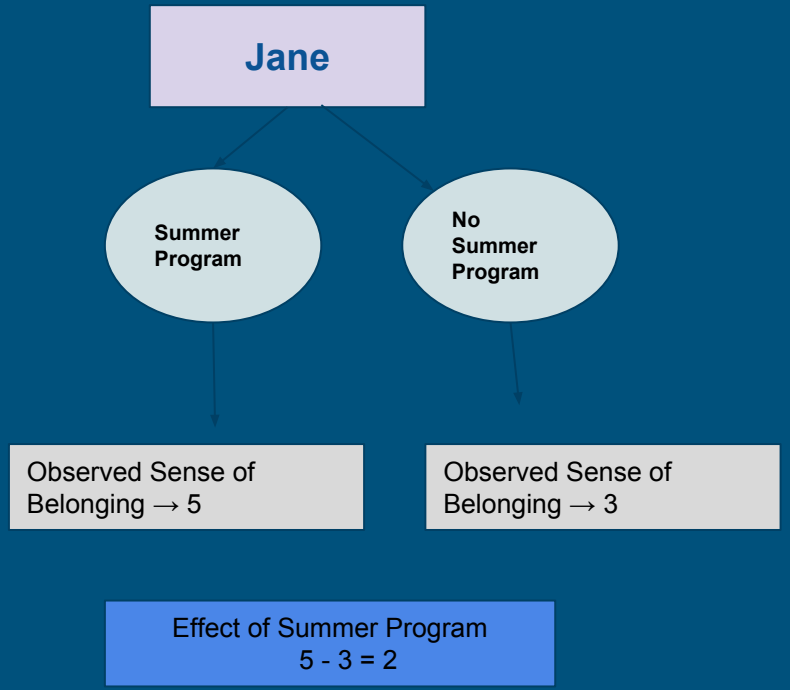
"Evaluation methods using an experimental design are best for determining project effectiveness. Thus, the project should use an experimental design under which participants--e.g., students, teachers, classrooms, or schools--are randomly assigned to participate in the project activities being evaluated or to a control group that does not participate in the project activities being evaluated." U. S. Department of Education, 2003.

The AEA Responds

“While we agree with the intent of ensuring that federally sponsored programs be "evaluated using scientifically based research . . . to determine the effectiveness of a project intervention," we do not agree that "evaluation methods using an experimental design are best for determining project effectiveness.” We believe that the constraints in the proposed priority would deny use of other needed, proven, and scientifically credible evaluation methods, resulting in fruitless expenditures on some large contracts while leaving other public programs unevaluated entirely.”
AEA, 2003, Emphasis mine

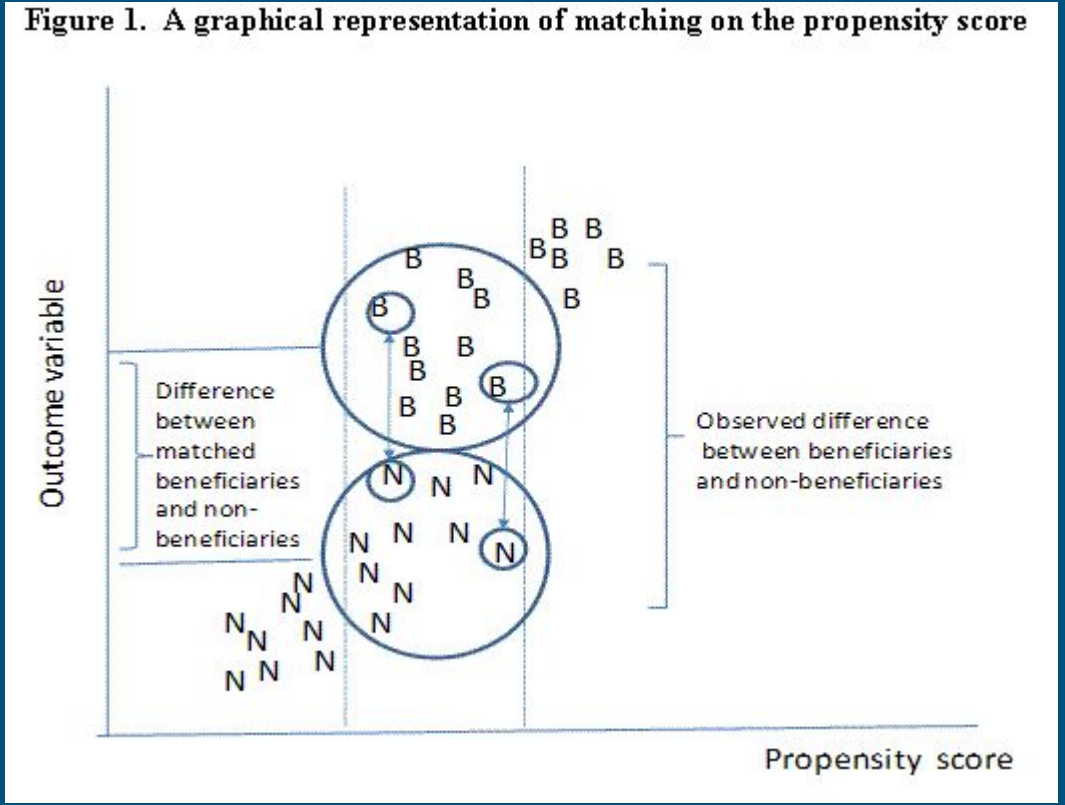
Rubin’s Causal Model

"The causal effect of a treatment on a single individual or unit of observation is the comparison (e.g. difference) between the value of the outcome if the unit is treated and the value of the outcome if the unit is not treated." (Angrist, Imbens, & Rubin, 1996)



Propensity Score Matching

- Estimate propensity score
  - A propensity score is the probability of treatment assignment based on a set of given covariates
- Decide on matching strategy
  - One-to-one matching
  - Greedy Matching
  - Many-to-one matching



Problem 1 - The Context

- Fellowship program
  - Financial and programming component
  - Competitive award with standard criteria
  - Program experienced changes in criteria in the 2008 AY
- What is the impact of program on doctoral outcomes?
  - Time to Degree
  - Time to Candidacy
- How is the impact different across divisions, schools/college, departments?

Simple Descriptives and Comparisons

Year	Number of Students Receiving Fellowships	Average TTC	Average TTD	% of Fellows Reaching Candidacy	% of Fellows Earning Doctorate
2004	112	2.91	5.89	85%	65%
2005	114	2.76	5.67	72%	42%
2006	98	2.64	5.17	79%	39%
2007	116	2.69	4.60	82%	20%
2008	135	2.51	3.35	76%	2%
2009	131	2.17		55%	N/A
2010	128	1.41		8%	N/A

Total N students in included cohorts = 7166

Total N students receiving fellowships = 834

Total N with time to Degree = 185

Total N with time-to-candidacy = 534

9% of total population are considered under-represented minorities

67% of students that receive fellowships are under-represented minorities

Comparisons Continued

Division / RMF Status	Cohort Size	Avg TTC	Avg TTD	% of Cohort Reaching Candidacy+	% of Cohort Earning Doctorate+
Biological and Health Sciences	**1514**	**2.00**	**4.90**	**74%**	**34%**
No Fellowship	1387	1.98	4.88	75%	35%
Fellowship	127	2.21	5.15	65%	22%
Physical Sciences and Engineerir	**3342**	**1.98**	**4.62**	**70%**	**35%**
No Fellowship	3039	1.93	4.57	72%	36%
Fellowship	303	2.58	5.34	56%	25%
Social Sciences	**1600**	**2.63**	**5.50**	**70%**	**25%**
No Fellowship	1298	2.61	5.48	71%	27%
Fellowship	302	2.75	5.63	67%	20%
Humanities	**652**	**2.49**	**5.56**	**70%**	**23%**
No Fellowship	557	2.48	5.49	70%	24%
Fellowship	95	2.60	6.03	73%	19%

Propensity-Score Matched Pairs

Fellowship

1 to Many Matching
Similar student:

- **School/College**
- **Gender**
- **URM/NonUrm**
- **Cohort**

Matched-Pair Comparisons

Simple T-test

Mean difference of about .42 years on time-to-degree (A bit more than a semester)

Mean difference of about .23 years on time-to-candidacy (a bit less than 1 semester)

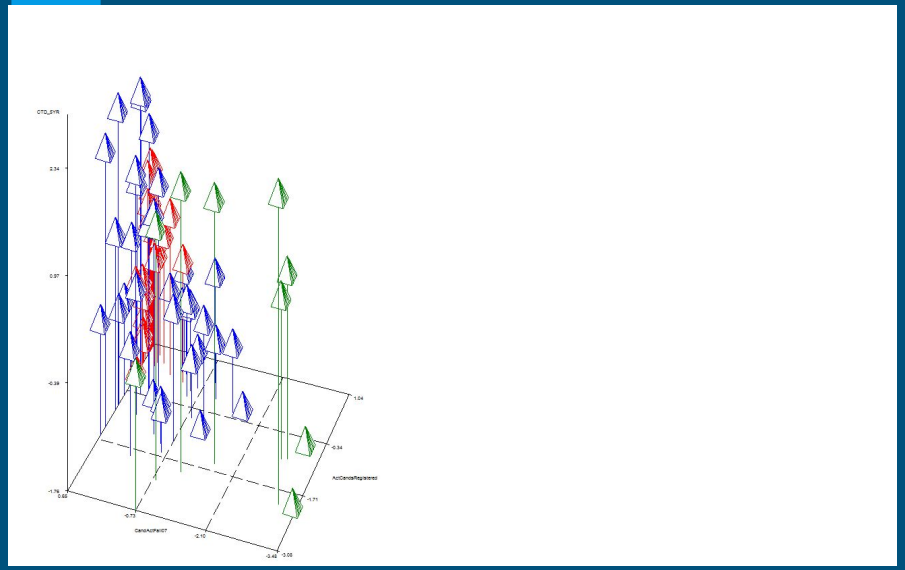
What’s next?

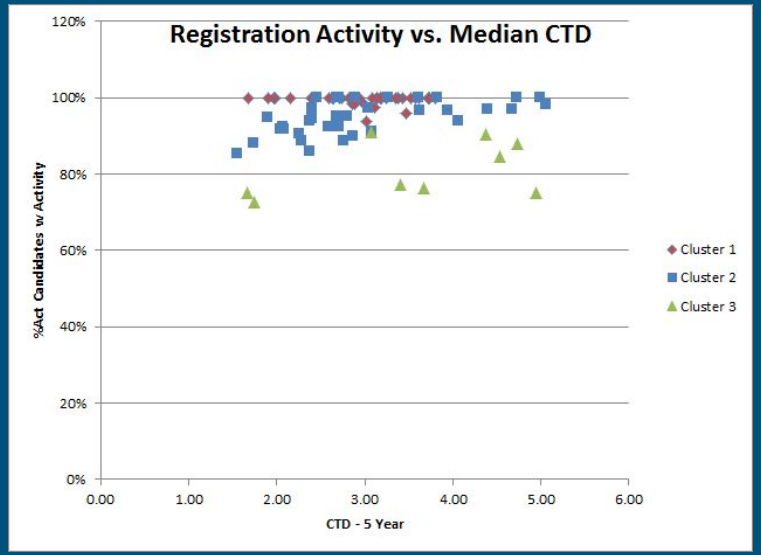
- Logistic regression with pairs to predict probability of completion
- More formal models to compare time-to-degree and time-to-candidacy across schools/colleges and controlling for other variables
- Exploring better way to match students - matching algorithm is a work in progress

Problem 2

- In 2010, a new policy went into effective requiring all doctoral students to register
- Five years later, we are asked to evaluate the impact the the policy on a number of indicators
- This presents a number of issues - namely how do we control for a number of concurrent interventions taking place?
- Is it possible to create a set of comparative groups that equalize the baseline to help us make better inferences?

Cluster Analysis





By equalizing across baseline variables we believe differentiate programs, we think we can create a set of groups. A baseline group to which the policy should have little or no effect. A “small effect” group that should have a smaller change, and a “high-impact” group that should see large changes due to the policy.

Problem 3

Suppose we want to assess the impact that highly-selective schools have on student outcomes? But we can’t randomly assign students to a school...and matching creates a problem because of unobserved, unmeasured characteristics (i.e. highly-motivated, parental involvement)

Enrollment in a Specialized High School and Future Outcomes

Figure 2
Exam School Eligibility and College Outcomes

Figure 2A: 4-year Enrollment

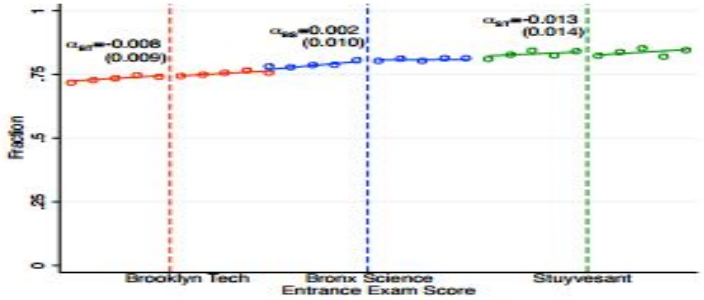


Figure 2B: 4-year Graduation

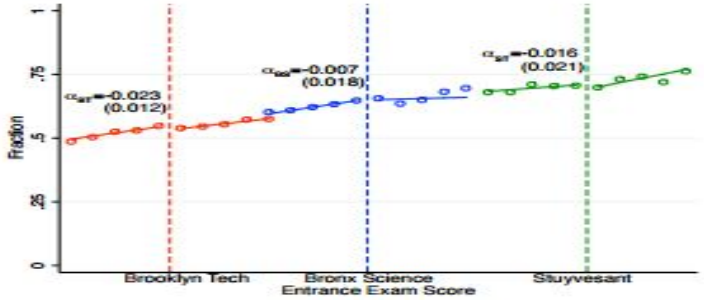


Figure 2C: Enrollment at SAT > 1200

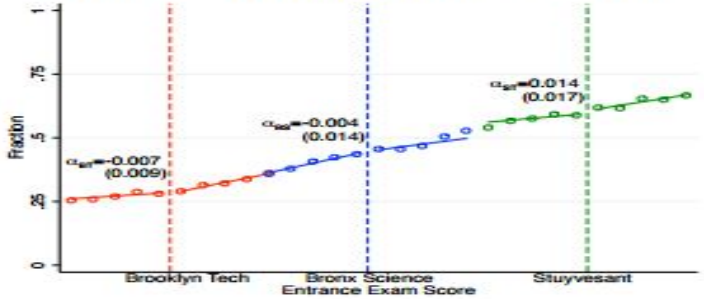
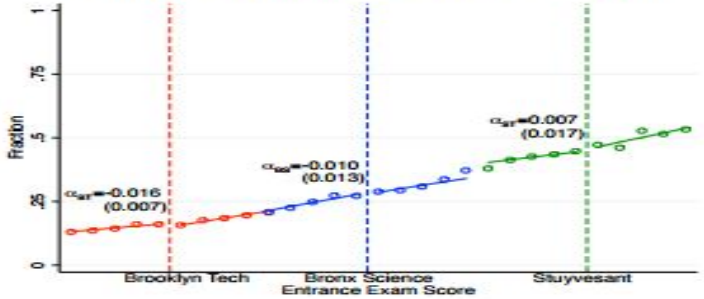


Figure 2D: Enrollment at SAT > 1300



Questions?

“Evaluation practice is not just applied social science methodology. Evaluation is a field that raises deeply interesting and challenging intellectual issues, a field that has developed a set of unique conceptualizations about how to deal with those issues. Those issues, and how we cope with them, are the stuff of evaluation theory, and they define who we are as a profession.“ Will Shadish, 1998