DAY 3
April 26, 2000

Session 1: Accounting for "No-Shows" in Randomized Experiments

Goals

This workshop session was designed to deepen participants' understanding of randomized experiments by focusing on how to deal with situations where not all persons who are randomly assigned to a program actually take part in it (and thus, become "no-shows"). Particular emphasis was placed on the conceptual issues involved in interpreting impact estimates from such studies, the statistical issues involved in producing these estimates, the assumptions necessary for their validity, and how to design experiments to reduce the magnitude of this problem.

Topics

- What are "no-shows", how do they differ from "attriters" and how do they affect program impact estimates?

- What is the impact of offering a program to potential participants? What is the impact of their participation in the program? What is the difference between these two questions?[1]

- Why can one estimate the impact of a program offer by comparing the subsequent outcomes of persons who were randomly assigned to a program or control group?

- In a randomized experiment, how can use estimates of the impact of a program offer plus a measure of the program "no-show" rate to estimate the impact of program participation? What is the intuition of this approach, what assumptions are necessary for it to work, and under what conditions are these assumptions likely to be met?

Reading

Bloom, Howard S. (1984) "Accounting for No-Shows in Experimental Evaluation Designs" ***Evaluation Review***, Vol. 8, No. 2, April, 225-246.

---

[1] These alternatives are often referred as the impact of "intention to treat" and the impact of "receiving treatment", respectively.

Accounting for "No-Shows" in Randomized Experiments[2]

MDRC
April 26, 2000

1. **What is the "no-show" (non-participant) problem?**

    - in social program evaluations
    - in medical clinical trials

2. **How do no-shows differ from "attriters"?**

    - no-shows do not receive program services but remain in the follow-up sample
    - attriters may or may not receive program services but do not remain in the follow-up sample

3. **How do no-shows affect program impact estimates?**

    - they dilute the treatment contrast
    - which reduces the treatment and control group response difference,
    - which, in turn, reduces statistical power

4. **What are the two types of impacts to estimate in the presence of no-shows?**

    - the impact of *offering* program services ("intent to treat")
    - the impact of *receiving* program services (treatment)

[2] From Bloom, Howard S. (1984) "Accounting for No-Shows in Experimental Evaluation Designs" *Evaluation Review*, Vol. 8, No. 2, April, pp. 225-246.

5. **How <u>should one not</u> estimate impacts in the presence of no-shows?**

- by comparing outcomes for service recipients with those for control group members (because you cannot identify the control group counterparts for service recipients)

6. **How <u>should one</u> estimate impacts in the presence of no-shows?**

- by comparing all program group members with all control group members (which is a purely experimental estimate of the impact of *offering* program services),
- by *adjusting* the program and control group outcome difference to reflect the no-show rate (discussed below)

7. **Derivation of the "semi-experimental" no-show adjustment**

The impact per experimental, I_E, equals the average of the impact per no-show, I_{NS}, and the impact per participant, I_P, weighted by the no-show rate, k. In other words:

$$\mathbf{I_E = kI_{NS} + (1-k)I_P}$$

If the impact per no-show is zero (or negligible) then

$$\mathbf{I_E = k(0) + (1-k)I_p}$$
$$\mathbf{= (1-K)I_p}$$

Therefore:

$$\mathbf{I_P = I_E/(1-k)}$$

8. Example of the no-show adjustment

Given that:

$\bar{Y}_E = 80$ pts $\quad n_E = 100 \quad k = 0.5$

$\bar{Y}_C = 75$ pts. $\quad n_C = 100$

then

$$\hat{I}_E = \bar{Y}_E - \bar{Y}_C = 80 - 75 = 5 \text{ pts.}$$

and

$$\hat{I}_P = \hat{I}_E/(1-k) = 5/(1 - 0.5) = 10 \text{ pts.}$$

9. **Intuition of the no-show adjustment**

Total score for experimentals	=	8,000 pts.
Total score for controls	=	7,500 pts.
Difference in total scores	=	500 pts.

a. Spreading the total difference across the 100 experimentals yields

500/100 = 5 points per experimental

b. Spreading the total difference across the 50 participants yields

500/50 = 10 points per participant

10. **Limitations of the no-show adjustment**

- Findings generalize to participants only
- Approach does not apply to mandatory programs (with sanctions)
- Approach does not apply to "partial treatments"

DAY 3
April 26, 2000

Session 2: Estimating Program Impacts on Student Performance Using "Short" Interrupted Time-Series

Goals

This session was intended to introduce workshop participants to the use of interrupted time-series analysis for estimating the impacts of comprehensive school reforms. In particular, the session was designed to introduce the basic logic of interrupted time-series analysis, illustrate how it could be used to estimate program impacts on student performance, present a simple estimation procedure for doing so, examine the statistical properties of this procedure, consider the assumptions that are necessary for the procedure to work, examine the implications of the procedure's data requirements for potential applications, and explore its strengths and weaknesses.

Topics

I. What is interrupted time-series analysis and how has it been used in other areas to measure program impacts?

II. How can this approach be applied to administrative data on standardized test scores to measure the impacts of educational programs on student performance?

III. How can a simple regression analysis be used to estimate program impacts using this approach?

IV. What are the strengths and weaknesses of the approach and how can it be expanded (using comparison series and multiple sites) to improve the reliability and validity of its impact estimates?

V. What are the statistical properties (especially the minimum detectable effects) of the approach given different configurations of baseline and follow-up data that are likely to be available in practice?

VI. How can one pool the results of such analyses across different schools that use different tests to measure student achievement?[3]

Reading

Bloom, Howard S. (1999) "Estimating Program Impacts on Student Achievement Using "Short" Interrupted Time-Series" (New York: Manpower Demonstration Research Corporation, August).

[3] The same test must be used over time at any given school, however.

Estimating Program Impacts on Student Performance Using "Short" Interrupted Time-Series[4]

MDRC
April, 2000

1. Introduction

- **The evaluation problem:** measuring the impacts of comprehensive school reform initiatives on student performance
- **Focus of the analysis:** educational excellence and equity
- **Need for the approach:** when random assignment of individuals, classes or schools is not possible
- **Background of the approach:** used widely in policy areas other than education

2. Approach

- **The data:** standardized test scores for a number of years before an initiative (the baseline period) and a number of years after the initiative is launched (the follow-up period)
- **The logic:** impacts are measured as deviations from past trends (Figure 1)

[4] From Bloom, Howard S. (1999) "Estimating Program Impacts on Student Achievement Using "Short" Interrupted Time-Series" (New York: Manpower Demonstration Corporation, August).

- **The regression model** (for a single program school with "cohort effects" but without individual covariates):

$$Y_i = A + B\,t_i + D_0\,F_{0i} + D_1\,F_{1i} + D_2\,F_{2i} + D_3\,F_{3i} + D_4\,F_{4I} + u_t + e_i$$

where:

Y_i = the test score for student i,
t_i = the test year for student i (ranging from - 5 through + 4 in Figure 1)
F_{0i} = 1 if student i took the test in follow-up year zero and 0 otherwise,
F_{1i} = 1 if student i took the test in follow-up year one and 0 otherwise,
F_{2i} = 1 if student i took the test in follow-up year two and 0 otherwise,
F_{3i} = 1 if student i took the test in follow-up year three and 0 otherwise,
F_{4i} = 1 if student i took the test in follow-up year four and 0 otherwise,
e_i = the random *individual difference* in the score for student i (which is independently and identically distributed across students in a year with a mean of zero and a variance of σ^2),
u_t = the random *annual cohort difference* in the mean score for year t (which is independently and identically distributed across years with a mean of zero and a variance of τ^2),
A and B = the intercept and slope of the baseline trend respectively, and
D_0, D_1, D_2, D_3 and D_4 = deviations from the baseline trend (impact estimates) for follow-up years 0, 1, 2, 3 and 4 respectively.

- **Strengths of the approach**
  - protection against maturation effects
  - protection against regression artifacts

- **Limitations of the approach**
  - susceptibility to history effects
  - susceptibility to selection bias
  - susceptibility to instrumentation effects

- **Extensions of the approach**
  - using comparison series
  - using multiple sites

3. Minimum Detectable Effect Size, MDES (with "cohort effects" and no individual covariates)

- **Formula** (Bloom, 1999, Equation 4, p.10)

$$MDES(\bar{D}_f) = \frac{2.5}{\sqrt{m}}\sqrt{1/n + r/(1-r)}\sqrt{1+\frac{1}{T}+\frac{(t_f-\bar{t})^2}{\sum_k (t_k-\bar{t})^2}}$$

- **Implications for a Single School with a small cohort effect (**r = 0.01)

Table 1

Minimum Detectable Effect Size for One School
By Follow-up Year and Baseline Period
(cohort effect r = 0.01)

Follow-up Year	Baseline Period		
	Four Years	Five Years	Six Years
	50 Students Each Year		
Zero	0.69	0.63	0.54
One	0.83	0.73	0.59
Two	1.00	0.83	0.66
Three	1.17	0.95	0.74
Four	1.35	1.07	0.82
	75 Students Each Year		
Zero	0.61	0.56	0.47
One	0.74	0.64	0.52
Two	0.88	0.74	0.58
Three	1.03	0.84	0.65
Four	1.19	0.95	0.72
	100 Students Each Year		
Zero	0.56	0.51	0.44
One	0.68	0.59	0.48
Two	0.82	0.68	0.54
Three	0.96	0.78	0.60
Four	1.10	0.88	0.67

- **Implications for Multiple Schools with varying cohort effects (r)**

Table 2

Minimum Detectable Effect Size
By Number of Program Schools,
Students Per Grade, and Cohort Effect (r)
(for follow-up year 2 with 5 baseline years)

Number of Program Schools	Students per Grade		
	50 Students	**75 Students**	**100 Students**
		r = **0.01**	
1 School	0.83	0.74	0.68
5 Schools	0.37	0.33	0.31
10 Schools	0.26	0.23	0.22
20 Schools	0.19	0.17	0.15
30 Schools	0.15	0.13	0.12
40 Schools	0.13	0.12	0.11
		r = **0.03**	
1 School	1.09	1.01	0.97
5 Schools	0.49	0.45	0.44
10 Schools	0.34	0.32	0.31
20 Schools	0.24	0.23	0.22
30 Schools	0.20	0.19	0.18
40 Schools	0.17	0.16	0.15
		r = **0.05**	
1 School	1.30	1.24	1.20
5 Schools	0.58	0.55	0.54
10 Schools	0.41	0.39	0.38
20 Schools	0.29	0.28	0.27
30 Schools	0.24	0.23	0.22
40 Schools	0.21	0.20	0.19

4. Further Issues

- **Using aggregate data**

    - **The issue:** data are often readily availabile,
    - **The good news:** point estimates are the same for aggregate and individual-level models
    - **The bad news:** some precision is lost due to the limited number of degrees of freedom

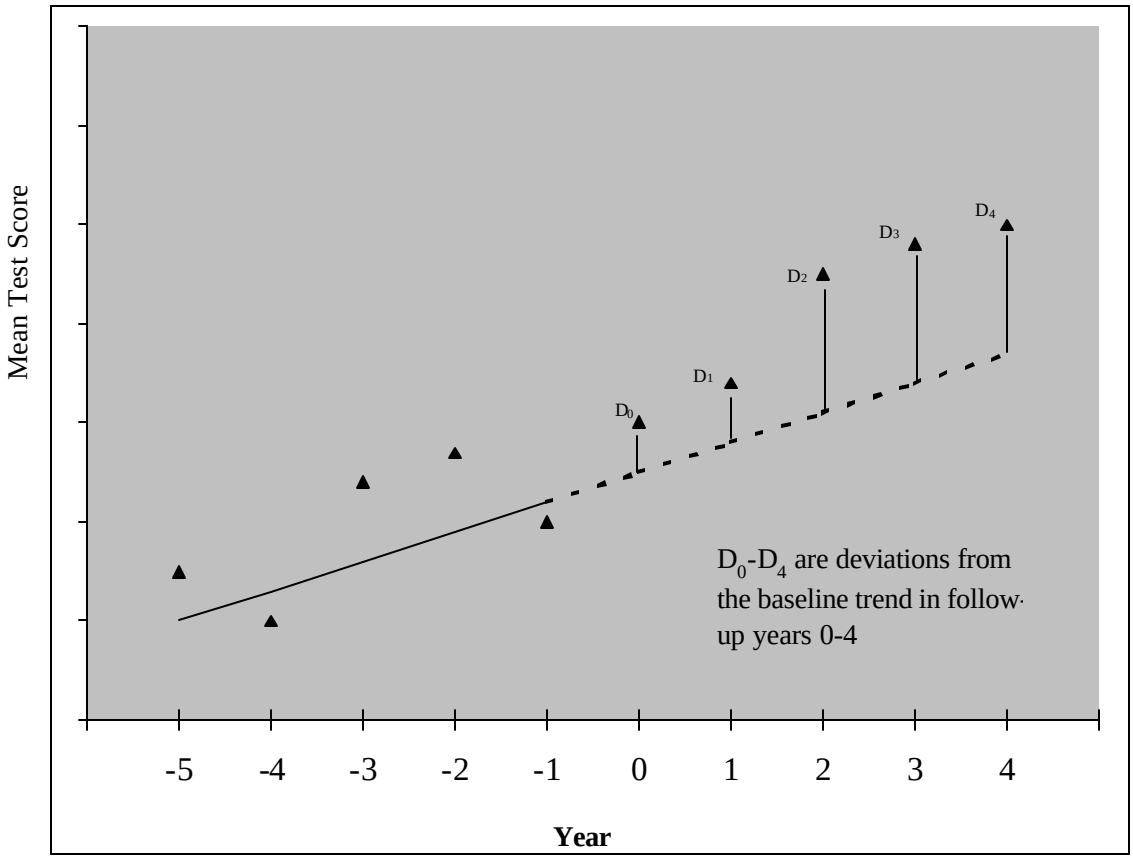
- **Pooling Findings Across Schools**

    - Requires same test over time for each school
    - But can accommodate different tests for different schools

- **Next Steps in Developing the Methodology**

    - Exploring different baseline trends
    - Combining interrupted time-series analysis with value-added analysis and hierarchical modeling

An Hypothetical Interrupted Time-Series Analysis of Impacts on Mean Test Scores



Note: Years -5 through -1 in the figure represent the pre-reform baseline period, and the solid line through the mean test scores for these years represents the estimated baseline trend. The dashed line in years O through 4 represents the extrapolation of this trend into the follow-up period and serves as the counterfactual for the impact analysis. Year 0 represents the launch year of the reform and years 1 through 4 represent subsequent follow-up years.

Day 3
April 26, 2000

Session 3: Theory of Change Evaluation

Goals

This session focused on the role of program theory in evaluation. Such theory can provide much-needed clarity about the inputs, processes, and desired outcomes of the initiative or program being examined. This is especially critical for evaluating "comprehensive community initiatives", like whole-school reforms. To illustrate the role that program theory can and should play in evaluation, The PES evaluation of Upward Bound was used as a case study.

Topics

I. Peter Rossi's "Metallic Laws of Program Evaluation," and their suggestion that many program evaluations show limited effects because the nature of the problem or its solution has been poorly conceptualized,

II. The special importance of theory in initiatives that seek to engage all community members, that seek to achieve effects in a variety of areas and at many different levels, and that are not readily definable,

III. The role of theory in guiding the development of measures and the collection of data,

IV. The role of theory in enabling evaluators to get "inside the black box" of a program to determine how it did or did not make a difference,

V. The need for information about the counterfactual for the program treatment, as well as for program outcomes,

Readings

Carol Hirschon Weiss, "Nothing as Practical As Good Theory: Exploring Theory -Based Evaluation for Comprehensive Community Initiatives for Children and Families." ***In New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts***, Vol. 1, James P. Connell et al., eds. (Washington, D.C.: The Aspen Institute, 1995).

James P. Connell, J. Lawrence Aber, Gary Walker, "How Do Urban Communities Affect Youth? Using Social Science Research to Inform the Design and Evaluation of Comprehensive Community Initiatives." In *Ibid.*

PETER ROSSI'S "METALLIC LAWS" OF PROGRAM EVALUATION

The "Iron Law": The expected value of any net impact assessment of any social program is zero.

The "Stainless Steel Law": The better designed the impact assessment of a social program, the more likely is the net effect to be zero.

The "Copper Law": The more social programs are designed to change individuals, the more likely the net impact of the program will be zero.

The "Plutonium Law": Program operators will explode when exposed to typical evaluation research findings.

WHY DO SO MANY EVALUATIONS HAVE SUCH LIMITED EFFECTS? (ACCORDING TO ROSSI)

1. Poor conceptualization of the problem.

2. Poor conceptualization of the solution.

3. Poor implementation.

COMPREHENSIVE COMMUNITY INITIATIVES (CCIs)

What they are

Efforts to promote positive change in individual, family, and community circumstances in disadvantaged neighborhoods by improving physical, social, and economic conditions.

Why the are hard to evaluate using more conventional methods

- They aren't discrete programs or initiatives with well-specified procedures.
- They seek to achieve change at many levels.
- They have multiple goals.
- They try to engage and affect everyone in the community.
- They are works in process and change over time.

HOW IS A THEORY OF CHANGE USEFUL?

- Planning and resource allocation
- Measurement and data collection
- Midcourse corrections
- Getting “inside the black box”

ESTABLISHING CAUSATION IN A THEORY OF CHANGE: SOME PROPOSITIONS

The larger the effect, the less rigorous the research design needs to be.

- The more we can rule out plausible alternative explanations, the more likely it is that the theory of change is valid.

- The more important the decision we need to make, the stronger our evidence should be.

- Causation is a less salient issue at the early stages of a theory of change than at the later stages.

- Nonetheless, the more we know about the counterfactual at every stage of the theory of change, the better we can "unpack" the elements of an intervention to understand what's going on.