DAY 1
December 6, 1999

Session 1: Basic Concepts

Goals

This session provided an introduction to and foundation for the workshop series. It was designed to identify the types of evaluation questions that would be addressed, help workshop participants understand what it means to measure the impacts of a program, introduce them to the basic building blocks of research designs for doing so, sensitize them to the main limitations of these designs, illustrate how regression analysis can be used to strengthen these designs, and outline the role that randomized experiments can play in this process.

Topics

- The types of evaluation questions to be addressed by the workshop series (specifically, how to assess program impacts and implementation),

- When to conduct a study of program impacts or implementation (and when not to do so),

- The logic of making causal inferences about program impacts (identifying outcomes, counterfactuals and impacts),

- The basic building blocks of research designs for estimating program impacts (by comparing outcomes across groups and/or time),

- Threats to the internal and external validity of program impact estimates (limitations on making and generalizing causal inferences),

- The use of regression analysis to estimate program impacts,

- Why and when to use randomized experiments to measure program impacts.

WORKSHOP ON EVALUATING PROGRAM IMPACTS AND IMPLEMENTATION

DAY 1: BASIC CONCEPTS

by

**Howard Bloom and
Janet Quint**

**Manpower Demonstration
Research Corporation**

for the

**Program Evaluation Service,
US Department of Education**

December 6, 1999

WORKSHOP CONTENTS

What types of evaluations will we cover?

- *Impact evaluations* of program or policies,

- *Implementation (process) evaluations* in the context of impact studies,

What types of evaluations will we not cover?

- *Needs assessments* – systematic analyses of the type, nature, and prevalence of a problem,

- *Diagnostic evaluations* – procedures to define and specify problems in ways that lead to the design of appropriate interventions,

- *Program monitoring* – assessments of whether programs are being run in conformity with specific operational objectives and plans,

- *Replication assessments* - analyses of a program's replication potential,

- *Cost-benefit analyses* – comparisons of the costs and impacts of a program, with both expressed in monetary terms,

- *Cost-effectiveness analyses* – analyses of the cost per unit impact of an intervention,

WHEN TO EVALUATE?

- when many people care about the problem, and the question is an *important* one,

- when the program model is reasonably *well specified*,

- when there is reason to think that the program can *produce desired outcomes*,

- when the program operates at a *large enough scale* to produce reliable estimates of effects,

- when the program is a *unique addition/contrast* to the array of services that would otherwise be in place (including alternative interventions),

- when program operators are *willing to cooperate* with the research requirements,

- when the findings of the evaluation are likely to *guide action*,

- when the need for answers is *urgent but not immediate*,

MEASURING PROGRAM IMPACTS

1. Perspectives

- Learning this material is like *peeling* an onion.
- You don't really understand it until you can *say* it.
- The key to getting an answer is framing the *question*.
- No single study can provide definitive answers. The best it can do is help to *accumulate knowledge*.
- There is no proof—only more or less *plausible arguments*.

2. Overview

- The logic of causal inference
- Basic impact study designs
- Common threats to the validity of impact estimates
- Regression-based impact estimates
- Randomized experiments

3. Outcomes, Counterfactuals and Impacts

- **Outcomes** = what happened *with* the program
- **Counterfactuals** = what would have happened *without* the program
- **Impacts** = what happened *because of* the program
- **Measures** = test scores, attendance rates, graduation rates, student engagement scales, teacher commitment scales, *etc.*

4. Basic Impact Study Designs

- **<u>One group before-after designs</u>**
  (based on group *changes*)

$$Y_{P1} \quad\quad P \quad\quad\quad Y_{P2}$$

$$\textbf{Impact Estimate} \; = \; Y_{P2} \, \text{-} \, Y_{P1}$$

- **<u>Two group after-only designs</u>**
  (based on group *differences*)

$$P \quad\quad\quad Y_{P2}$$

$$Y_{C2}$$

$$\textbf{Impact Estimate} \; = \; Y_{P2} \, \text{-} \, Y_{C2}$$

- **<u>Two group before-after combination designs</u>**
  (based on *differences* in group *changes* or
  changes in group *differences)*

$$Y_{P1} \qquad P \qquad Y_{P2}$$

$$Y_{C1} \qquad\qquad Y_{C2}$$

$$\textbf{Impact Estimate} \ = \ (Y_{P2} - Y_{P1}) \ - \ (Y_{C2} - Y_{C1})$$

$$= (Y_{P2} - Y_{C2}) \ - \ (Y_{P1} - Y_{C1})$$

5. Threats to Internal Validity
(alternative explanations for what was observed)

- *Maturation* = underlying systematic changes that can occur over time

- *Selection* = initial differences that can exist between the groups being compared

- *History* = specific events or activities other than the program that can affect outcomes

- *Testing* = changes in the meaning of outcome measures that can occur if respondents adapt to the measurement process

- *Instrumentation* = changes in the meaning of outcome measures that can be caused by changes in how they are administered or reported

- *Regression artifacts* = changes that can occur if the starting point is unusual

- *Attrition* = differences in the groups being compared that occur if follow-up data is missing for some of their members

6. Threats to External Validity
(limits to generalizability)

- *Selection/treatment interactions* = differences in impacts for different types of subjects

- *Setting/treatment interactions* = differences in impacts for different types of settings

- *History/treatment interactions* = differences in impacts for different times or periods

7. Using Multiple Regression Analysis to Estimate Program Impacts

The Basic Model

$$Y_i = a + b_0 P_i + b_1 X_{1i} + b_2 X_{2I} + \cdots + e_i$$

or

$$Y_i = a + b_0 P_i + \mathbf{S} b_j X_{ji} + e_i$$

where:

$Y_i =$ the outcome for student i,

$P_i =$ one for students in the program group and zero otherwise,

$X_{ji} =$ background characteristic j for student i (which could include a pre-test score),

$b_0 =$ the program impact,

$b_j =$ the regression coefficient for background characteristic j,

$a =$ the intercept of the regression,

$e_i =$ the random error term for student i.

$s =$ the standard deviation of Y_i for the control group,

$R^2 =$ the percentage of the variation in Y_i "explained" by P_i and the X_js,

Hypothetical Example #1 (without pre-test)

POST-TEST $= 400 + 60$ **PROGRAM** $- 25$ **MALE** $+ 30$ **WHITE**
        (20) (25)   (20)   (15)

           $- 25$ **FREE-LUNCH**
           (10)

$s = 100$
$R^2 = 0.15$
standard errors are in parentheses below each parameter estimate

Hypothetical Example #2 (with pre-test)

POST-TEST $= 125 + 35$ **PROGRAM** $- 5$ **MALE** $+ 15$ **WHITE**
        (15)   (5)   (15)   (10)

           $- 15$ **FREE-LUNCH** $+ 0.7$ **PRE-TEST**
          (7)   (0.1)

$s = 100$
$R^2 = 0.55$
standard errors are in parentheses below each parameter estimate

8. Using Random Assignment to Measure Program Impacts

- What is random assignment?

- What does random assignment accomplish?

- What is the role of regression analysis with random assignment?

DAY 1
December 6, 1999

Session 2: Evaluating a Tutoring Program for Elementary School Students: A Mini-Case

Goals

This session, which centered on a hypothetical intervention program, asked workshop participants to apply some of the concepts developed during Session 1. First, it asked participants to consider the evidence they would need to determine whether an evaluation of the program was warranted and to design such an evaluation. Then, because random assignment of research subjects into experimental and control groups is widely recognized to yield the most reliable estimates of program impacts but is often believed to be infeasible, attendees took part in a role play to better understand the objections to random assignment and how to counter them

Topics

- Criteria for determining whether an intervention is worth a rigorous test (solid underlying theory, prior evidence supporting the intervention, evidence on the strength of implementation),

- Choices related to research designs (principal questions, measures, selection of an appropriate evaluation design, data sources),

- The usefulness of implementation analysis for understanding program impacts or the lack thereof,

- Assessing the feasibility of a random assignment experiment,

- Arguments that can be used to persuade teachers and parents of the usefulness of random assignment to answer important questions,

Readings

Gueron, Judith M. (1999) "The Politics of Random Assignment: Implementing Studies and Impacting Policy." (New York, Manpower Demonstration Research Corporation)

Sherwood, Kay E. and Fred Doolittle, (2000) "What's Behind the Impacts: Doing

Implementation Research in the Context of Program Impact Studies." (New York, Manpower Demonstration Corporation, prepared for the Institute for Research on Poverty).

Evaluating a Tutoring Program
for Elementary School Students:
A Mini-case

by
Janet Quint
Howard Bloom
MDRC

for
The Program Evaluation Service
of the U.S. Department of Education

December 6, 1999

The Setting

A branch of a state university is located in a major metropolitan center with a large low-income population. The student body in the school district is 40 percent African-American and 25 percent Latino, with substantial numbers of newly arrived immigrants from Russia, Kosovo, and Indonesia. Forty-eight percent of students in the district qualify for free or reduced-price lunches, and 60 percent test below grade level in reading.

A professor at the university decides to offer a new course in which university students receive 3 credits for tutoring elementary school students in reading for four hours a week. (Other course requirements include assigned readings, weekly meetings to learn about reading instruction and to discuss experiences in the field, plus a term paper that summarizes what students have learned) The professor arranges for the university to provide a small subsidy for travel expenses, books and other materials.

During the first year, 50 students enroll in the course. They work in two of the district's 43 K-5 schools, where they receive a warm reception from school administrators and teachers. Because tutoring sessions take place after regular school hours, parents must give approval for their children to participate. Some university students work with one child for all four hours a week; others work with two children for two hours each. Although the effort seeks to focus on low-achieving students, it is open to all children who are interested and able to attend. At the end of the course, teachers in the schools report that students who received tutoring seemed more eager to learn and more cooperative than previously, but no tests were administered to measure their reading gains.

The Proposal

The professor has an idea that he hopes will bring him fame, acclaim, and articles in scholarly journals for many years to come: he will use Reading Excellence Act funds to expand the course and make it a requirement for the university's 400 education majors. He submits a proposal for funding to the U.S. Department of Education, asserting that project participation will result in increased reading scores for low-achieving students in the district.

ED is trying out a new procedure for selecting projects for funding under the REA. First, ED staff will conduct a preliminary screening of proposals to identify those which sound promising. Once projects have passed this screen, ED staff will provide guidance in helping to shape a rigorous evaluation of each.

The professor's proposal is assigned to you. You find his concept to be intriguing, but his proposal is silent about the project's theoretical underpinnings and contains virtually no information about how things went during the pilot phase.

Your Response

Please think about the following questions and discuss them with others.

1. ***What do you need to know in order to determine whether the project is worth funding as a demonstration?*** If you were writing a letter asking the professor to clarify his proposal, what questions would you ask him? (see "What's Behind the Impacts: Doing Implementation Research in the Context of Program Impact Studies" by Kay E. Sherwood and Fred Doolittle for some helpful suggestions.)

2. *After receiving satisfactory answers to the first set of questions, you decide to move forward with the project.* ***How would you design an evaluation of the project's impacts?*** Specifically:

   What questions would you seek to answer?
   What outcomes would you try to measure?
   What research design would you employ?
   What data sources would you use?
   What strengths and weaknesses of your approach can you identify?

3. ***How would you design an evaluation of the project's implementation*** that would be helpful in understanding its impacts, or lack thereof? (Again, you may want to consult the Sherwood and Doolittle paper). Specifically:

    What questions would you seek to answer?
    What research methods would you employ?
    What data sources would you use?
    What strengths and weaknesses of your approach can you identify?

4. *You decide that a random assignment experiment would yield the strongest possible evidence.* ***How would you determine whether or not it is feasible to mount such an experiment?*** (see "The Politics of Random Assignment: Implementing Studies and Impacting Policy" by Judith M. Gueron, for helpful suggestions).