DAY 2
February 3, 2000

Session 1: Sample Size and Allocation for Randomized Experiments[1]

Goals

This session was intended to help workshop participants better understand how the size and allocation of a one's sample, plus other factors, such as the presence of longitudinal outcome data, can affect the reliability of program impact estimates. The session was designed to introduce the concept of statistical power, consider how it affects program impact estimates, present a simple way to assess the power of alternative evaluation designs, examine the main factors that affect their power, and explore ways to increase the power of future program impact studies.

Topics

- Definitions and examples of statistical power and statistical significance in the context of hypothesis tests about program impacts,
- Using the metric of "minimum detectable effects" to assess the statistical power of program impact studies,
- How sample size and allocation affect the minimum detectable effects of randomized experiments,
- How longitudinal outcome data and one-tail significance tests can reduce the minimum detectable effects of randomized experiments,
- Implications for the minimum detectable effects of non-experimental comparison group designs,
- The relationship between minimum detectable effects and minimum detectable effect size.

Readings

Bloom, Howard S. (1995) "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs" ***Evaluation Review***, Vol. 19, No. 5, October, 547-556.

Lipsey, Mark W. (1990) "Effect Size: The Problematic Parameter", from ***Design Sensitivity: Statistical Power for Experimental Research*** (Newbury Park, CA: Sage Publications) 47-56.

---

[1] Although the session focused mainly on randomized experiments, implications for non-experimental comparison group designs were considered as well.

Sample Size and Allocation for Randomized Experiments

MDRC Program Evaluation Workshop
for The Program Evaluation Service
of the US Department of Education

DAY 2
February 3, 2000

1. STATISTICAL HYPOTHESIS TESTING

- **Null and Alternative Hypotheses (H_0 and H_1)**

  H_0 = Not guilty
  H_1 = Guilty

  H_0 = No child abuse
  H_1 = Child abuse

  H_0 = No health benefit
  H_1 = Health benefit

  H_0 = No positive impact
  H_1 = Positive impact

- **<u>One-sided and Two-sided Alternative Hypotheses</u>**

  $\mathbf{H_0}$ = No impact
  $\mathbf{H_1}$ = Impact

  $\mathbf{H_0}$ = No positive impact
  $\mathbf{H_1}$ = Positive impact

  $\mathbf{H_0}$ = No negative impact
  $\mathbf{H_1}$ = Negative impact

- **<u>Type I and Type II Errors</u> (*see Figure 1*)**

  Type I Error = rejecting a true null hypothesis

  Type II Error = not accepting a true alternative hypothesis

  What do these errors mean in specific contexts?

  What do these errors imply (cost) in specific contexts?

  What does this suggest about structuring statistical hypothesis tests?

- **<u>Statistical Significance and Statistical Power</u>**

  - **Statistical significance (** α **)** = an indirect measure of how "real" an impact estimate is (usually assessed at the 0.05 level)

- **Statistical power (1 -** B) = an indirect measure of an evaluation design's ability to detect "true" impacts (usually assessed at the 80 percent level for a specified "true" impact)

- **NOTE:** One can assess the statistical power of a proposed evaluation design (*ex ante*) and one can assess the statistical power of a completed evaluation study (*ex post*).

2. FACTORS THAT INFLUENCE STATISTICAL POWER

- Population diversity
- Measurement reliability
- Sample size
- Sample allocation
- Background noise reduction
- **NOTE**: The standard error of an impact estimate reflects all of these factors.

3. REPRESENTING STATISTICAL POWER THROUGH THE MINIMUM DETECTABLE EFFECT (MDE)

- **<u>Intuition</u>**

  MDE = the smallest true impact that a specific design has a "good" chance of detecting

Examples

MDE = a 100 point increase on the SAT,

a $1,000 increase in earnings, or

a 10 percentage point increase in the graduation rate.

- **<u>Definition</u>**

  MDE = the smallest effect that, if true, has an X percent chance of producing an impact estimate that is statistically significant at the Y level

  Examples

  X = 80 percent power
  Y = 0.05 significance
  hypothesis test = one-sided (or two-sided)

- **<u>Calculation</u> (*See Table 1*)**

  MDE = Z times the standard error (**SE**) of the impact estimate

  Examples

  significance = 0.05
  power = 80 %

  MDE = 2.49***SE** (one-sided test)
  = 2.49*10 = 25 points

  MDE = 2.80***SE** (two-sided test)
  = 2.80*10 = 28 points

4. COMPUTING MINIMUM DETECTABLE EFFECTS FOR TWO-GROUP RANDOMIZED EXPERIMENTS

- **<u>The Basic Impact Regression</u> (for continuous and binary outcomes)**

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

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 (which could be a pre-test score) for student i,
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, and
R^2 = the percentage of the variation in Y_i "explained" by P_i and the X_js,

- **The Minimum Detectable Effect for Continuous Outcomes, MDE_C**

$$MDE_C = Z s \sqrt{\frac{(1-R^2)}{P(1-P)n}}$$

where:

Z = a multiplier which converts the standard error of an impact estimator to its corresponding minimum detectable effect,

σ = the standard deviation of the continuous outcome,

R^2 = the explanatory power of the impact regression,

P = the proportion of sample members randomly assigned to the program group, and

n = the total number of sample members.

Examples

σ = 100 points P = 0.5 n = 400 Z = 2.49

R^2 = 0 (no covariates)
MDE_C = 25 points

R^2 = 0.45 (with pretest)
MDE_C = 18.5 points

- **<u>The Minimum Detectable Effect for Binary Outcomes, MDE_B</u>**

$$MDE_B = Z\sqrt{p(1-p)}\sqrt{\frac{(1-R^2)}{P(1-P)n}}$$

where:

Z = a multiplier which converts the standard error of an impact estimator to its corresponding minimum detectable effect,

π = the proportion of the study population with a successful outcome,

R^2 = the explanatory power of the impact regression,

P = the proportion of sample members randomly assigned to the program group, and

n = the total number of sample members.

Examples

$\pi = 0.6$ P = 0.5 n = 400 Z = 2.49

$R^2 = 0$ (no covariates) **$MDE_B = 0.12$**

$R^2 = 0.45$ (with pretest) **$MDE_B = 0.09$**

5. FACTORS THAT INFLUENCE MINIMUM DETECTABLE EFFECTS

- **The Role of Sample Size (n)**

  What happens to the MDE if you double the sample? if you quadruple the sample?

- **The Role of Background Noise Reduction (R^2)**

  What happens to the MDE if R^2 is zero? if it is one?

  What are the implications for R^2 of the findings in ***Table 2?***

  1. More recent pre-tests reduce noise by more than less recent pre-tests.

  2. Nevertheless, less recent pre-tests reduce noise appreciably.

  3. Adding data for a less recent pre-test to that for a more recent pre-test does not further reduce noise appreciably.

  4. Sixth grade tests are more predictable than third grade tests. Hence, pre-tests provide more noise reduction for sixth graders than for third graders.
  5. Data for a recent pre-test (with an R^2 between 0.45 and 0.55) can reduce the MDE by as much as *doubling* the sample.

What are some other important empirical questions to explore about R^2?

1. What additional noise reduction can be obtained from data on individual background characteristics?

2. How do the preceding findings vary across grades, types of tests and types of students?

- **<u>The Role of Sample Allocation (P)</u>**

1. What is the *maximum* value of P(1-P) and thus, what sample allocation *minimizes* the MDE, other things being equal?

2. What do the findings in ***Table 3*** suggest about the role of sample allocation? How does this affect the way that you think about it?

- **<u>The Role of Outcome Variability [</u>**s **<u>or</u>** $p(1-p)$**<u>]</u>**

1. How does the reliability of one's outcome measure affect the MDE? What can be done to improve this reliability?

2. How does the diversity of one's population affect the MDE? What evaluation design tradeoffs does this suggest?

3. For what value of π is the MDE for a binary outcome largest? smallest?

4. How does the MDE change with π between these extremes?

5. What does this suggest for predicting the MDE of a binary outcome when designing an experiment?

- **<u>The Role of One-sided vs. Two-sided Tests</u>**

  Recall that for 0.05 significance and 80 % power:

  MDE = 2.49***SE** for one-sided tests
  MDE = 2.80***SE** for two-sided tests

  Therefore, one-sided tests provide more statistical power.

6. IMPLICATIONS FOR SIMPLE NON-EXPERIMENTAL COMPARISON GROUP ANALYSES

- The data and the impact regression are the same.

- The impact estimator no longer has the protection of random assignment and thus, is subject to selection bias.
- The standard error of the impact estimator is larger than its experimental counterpart because of the correlation between the program variable, P_i, and the background characteristics, X_{ji}. *(This is another way of saying that the background characteristics of program and control group members are different.)*

- The R_A^2 of the following auxiliary regression summarizes the correlation between P_i and the X_{ji}.

$$P_i = \alpha + \Sigma\beta X_{ji} + u_i$$

- Expression for the minimum detectable effect of a non-experimental comparison for a continuous outcome

$$MDE_C = Z s \sqrt{\frac{(1-R^2)}{P(1-P)n(1-R_A^2)}}$$

- **Therefore**, the minimum detectable effect of a non-experimental comparison is related to its experimental counterpart as follows:

$$\mathbf{MDE}_{\text{NONEXPERIMENTAL}} = \sqrt{\frac{1}{1-R_A^2}}\ \mathbf{MDE}_{\text{EXPERIMENTAL}}$$

- **Example**

  $R_A^2 = 0.2$

$$\mathbf{MDE}_{\text{NONEXPERIMENTAL}} = \sqrt{\frac{1}{1-0.2}}\ \mathbf{MDE}_{\text{EXPERIMENTAL}}$$

$$\mathbf{MDE}_{\text{NONEXPERIMENTAL}} = 1.12 * \mathbf{MDE}_{\text{EXPERIMENTAL}}$$

7. MINIMUM DETECTABLE "EFFECT SIZE" (MDES)

- Standardizing impacts as effect size (**ES**) by dividing each by the standard deviation of its corresponding outcome measure.

Example

Impact = 5 points
Standard deviation = 20 points

Effect Size = **ES** = 5/20
= 0.25 standard deviations

- Gauging the magnitude of an effect size (*How big is big?)*

- Cohen's Canon[2]

ES = 0.20 = *small*
0.50 = *medium*
0.80 = *large*

- Lipsey's Litany[3]

Distribution of Mean Effect Size

Range	Values	Midpoint
Small (bottom 3rd)	0.00 - 0.32	0.15
Medium (middle 3rd)	0.33 - 0.55	0.45
Large (top 3rd)	0.56 - 1.20	0.90

NOTE: Findings are based on 102 selected mean effect size estimates from 186 meta-analyses of 6,700 studies involving 800,000 subjects.

[2] Cohen, J. (1977) ***Statistical Power Analysis for the Behavioral Sciences*** (rev. ed.) (New York: Academic Press).
[3] Lipsey, Mark W. (1990) ***Design Sensitivity: Statistical Power for Experimental Research*** (Newbury Park, CA: Sage Publications), Table 3.5, p. 56.

Table 1

Multipliers to Convert the Standard Error of an Impact Estimate to its Corresponding Minimum Detectable Effect

	Significance Level		
Statistical Power	**0.10**	**0.05**	**0.01**
For One-sided Tests			
90 percent	2.56	2.93	3.61
80 percent	2.12	2.49	3.17
70 percent	1.80	2.17	2.85
For Two-sided Tests			
90 percent	2.93	3.24	3.86
80 percent	2.49	2.80	3.42
70 percent	2.17	2.48	3.10

SOURCE: Bloom, Howard S. (1995) "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs", ***Evaluation Review***, Vol. 19, No. 5, Table 1, p. 550.

Table 2

R^2 for Post-tests Regressed on Pre-tests Using Data for 25 Elementary Schools in Rochester, New York for 1992 and 1991

	Controlling for Pre-test in:		
6th Grade Post-test	**5th Grade**	**4th Grade**	**Both Grades**
1992 Math	0.42	0.36	0.47
1991 Math	0.47	0.35	0.50
1992 Reading	0.61	0.37	0.65
1991 Reading	0.56	0.46	0.61
	Controlling for Pre-test in:		
3rd Grade Post-test	**2nd Grade**	**1st Grade**	**Both Grades**
1992 Math	0.32	0.21	0.38
1991 Math	0.30	0.24	0.36
1992 Reading	0.55	0.32	0.57
1991 Reading	0.48	0.31	0.50

SOURCE: MDRC analyses of Pupil Evaluation Program (PEP) test scores for all third-graders and sixth-graders from the 25 Rochester elementary schools that had both grades in 1991 and 1992. The PEP is an annual norm-referenced test administered by the State of New York.

Table 3

The Effect of Sample Allocation on the Minimum Detectable Effect (MDE)

Program/Control Ratio	$MDE/MDE_{optimal}$ Ratio	Example One	Example Two
50/50 (optimal)	1.00	$1,000	5.0 %
60/40	1.02	1,020	5.1
70/30	1.09	1,090	5.5
80/20	1.25	1,250	6.3
90/10	1.67	1,670	8.4

NOTE: Reversing the program/control group ratio—for example, from 80/20 to 20/80—does not affect the minimum detectable effect.

DAY 2
February 3, 2000

Session 2: Interpreting An Evaluation Study: The MDRC Career Academies Report

Goals

This session provided participants with a list of questions that readers of both implementation and impact evaluation reports should ask themselves in order to make informed judgments about the reliability, generalizability, and usefulness of the findings. The most recent report from MDRC's Career Academies evaluation, which employs a random assignment research design to study program impacts, was then used as a case study to demonstrate how the questions can be used to assess a particular effort.

Topics

I. The questions that are addressed by the study (and those that are not addressed),

II. The study's policy and research contexts,

III. The components, duration, target group, and counterfactual for the treatment,

IV. The study's design,

V. The nature of the sample and its subgroups,

VI. Measures and data collection strategies,

VII. Implications and limitations of the study,

Readings

Kemple, James J., and Jason C. Snipes. (2000). ***Career Academies: Impacts on Students' Engagement and Performance in High School: Executive Summary***. (New York: MDRC).

POINTS TO ADDRESS WHEN SUMMARIZING AND INTERPRETING AN EVALUATION STUDY

Study Setting

1. What evaluation questions does the study address?

What questions related to program *impacts* are addressed?
What questions related to program *implementation* are addressed?
What questions are *not* addressed?

2. What is the policy context of the study?

What policy problem is being addressed?
For whom is this a problem?
What, if any, specific proposals to address the problem are being currently considered? By whom?

3. What is the research context of the study?

What research has been done on the topic?
What important knowledge gaps remain?
How does the present study help to fill these gaps?

Program Treatment

4. What are the components of the treatment?

What is the rationale for these components?
What specific problems are they designed to address - i.e., how is the treatment supposed to produce its intended effects?

5. What is the target group for the treatment?

What is the rationale for selecting this group?
If there is no special target group, why is this the case?

6. What is the planned treatment duration?

7. What is the counterfactual for the treatment?

Evaluation Methods Used

8. What is the basic design of the study?

Is it quantitative, qualitative, or both?
If quantitative, what is the research design used (experiment, time series, comparison group, etc.)?
What is the duration of the study, and what explains this duration?

9. How was the study sample drawn?

To what extent is the study sample representative of all people receiving the treatment?
How are subgroups created within the sample?

10. How were the study data obtained?

What kinds of data are collected, and how?

11. What measures are used by the study?

What are the key *outcome* variables, and how are they measured? Are different outcomes examined at different points in time?
What are the key *treatment* variables, and how are they measured?
What *other* variables are included, and how are they measured?

Key Findings

12. What are the main findings of the study?

What are the main findings about implementation?
What are the main findings about impacts?
What are the main findings about subgroups?
What are the main findings about the relationship between implementation and impacts?

Interpretation of the Findings

13. How do the findings of the study add to, reinforce, or contradict findings from previous research?

14. What are the main limitations of the study, and how do they affect your interpretation of its findings?

What problems are associated with the measures used?
What are the key threats to the internal validity of the study?
What are the key threats to the external validity of the study?

15. What are the study's key implications?

What are the implications for policymakers?
What are the implications for program operators?
What are the implications for clients?

16. What questions remain for researchers to address?