DAY 4
September 15, 2000

Session 1: A Quasi-Experimental Strategy for Measuring the Impacts of Whole School Reforms

Goals

The purpose of this session was to introduce a comprehensive strategy that is being developed by MDRC to measure the impacts of whole school reforms. The strategy will combine interrupted time-series analysis with value-added analysis. The session described each approach, explored how the weaknesses of each could be offset by specific strengths of the other (when they are used together), illustrated how different versions of the approach are being used for MDRC evaluations of four major whole school reforms, and described how the approach could be used to measure program impacts on average student test scores, the variation in test scores and the full distribution of scores.

Topics

- A review of interrupted time-series analyses (which was introduced in a previous session) and a description of how MDRC is using this approach for an exploratory evaluation of the Accelerated Schools Program.

- How to use interrupted time-series analysis to measure program impacts on mean test scores, the standard deviation of scores and the full distribution of scores.

- How to pool impact estimates across different schools and how this process varies depending on the population to which one is trying to generalize (infer) impact findings.

- A quick look at value-added analysis, its conceptual foundation, how it is used to measure the impacts of educational programs, and its strengths and weaknesses.

- How value-added analysis can be used together with interrupted time-series analysis to measure program impacts, and how the strengths of each help to offset the primary weakness of the other.

- How different combinations of interrupted time-series analysis and value added analysis are being used or will be used for MDRC evaluations of four major whole school reforms: Accelerated Schools, Project GRAD, Talent Development Schools and First Things First.

A Quasi-Experimental Strategy for Measuring the Impacts of Whole School Reforms

Howard S. Bloom
MDRC
June 21, 2000

1. Development of the Approach at MDRC

- The Accelerated Schools Study
- The Project GRAD-Newark Study
- The First Things First Study
- The Talent Development Schools Study

2. Stage #1: Accelerated Schools

- Background of the study
- Design of the impact analysis

- Retrospective interrupted time-series analysis

2. (continued)

- 8 mature Accelerated elementary schools from different parts of the country,

- 10 years of consistent individual-level test data (5 baseline years and 5 follow-up years) plus some demographics,

- minimal external changes to each school or its student population

- ***proposed approach*** to estimating program impacts = *using the baseline <u>trend</u> to project the counterfactual* (see Figure 1)

- ***actual approach*** to estimating program impacts = *using the baseline <u>mean</u> to project the counterfactual*

- ***pooling findings across schools*** (which depends on the generalization of interest)
    - ***option #1:*** generalize to the population of schools in one's study *(a fixed-effect inference)*

- ***option #2:*** generalize to a broader population of schools *(a random-effects inference)*

3. Stage #2: Project GRAD-Newark

- Working with one large school district that has automated student data made a comparison series design *feasible*

- Having a test change during the follow-up period made a comparison series design *necessary*

- Having individual student pre-test data made a value-added analysis *possible*

4. Stage #3: First Things First, Talent Development Schools and the PES/Sloan Methodology Study

- First Things First

    - Focusing exclusively on secondary schools

    - Dealing with program-induced “compositional shifts”

    - Adding a “Theory of Change”

- Talent Development Schools

    - Extending the reach of the combined strategy

- The PES/Sloan Methodology Study

    - Combining interrupted time-series, value-added and hierarchical modeling

    - Using the data from Project GRAD-Newark

5. A Quick Look at Value-Added Analysis

- logic of the approach

VALUE-ADDED = OUTPUT – INPUT

- application of the approach to measuring student achievement

VALUE-ADDED = POST-TEST – PRE-TEST

- basic analysis (see Figure 2)
- regression specification

$$\mathbf{Y}_{\mathbf{ijt}} \mathbf{=} a + B_0\mathbf{P}_{\mathbf{ij}} \mathbf{+} B_1\mathbf{Y}_{\mathbf{ij(t\text{-}k)}}\ B_2\mathbf{X}_{\mathbf{ij}} \mathbf{+} \mathbf{e}_{\mathbf{j}} \mathbf{+} e_{\mathbf{ij}} \qquad (1)$$

- ***key limitation = selection/maturation bias*** (program and comparison students may be on different initial growth paths)

- ***an example*** from the evaluation of employment and training programs of how bad selection/maturation bias can be (see Figures 3 and 4)[1]

- addressing the limitations of value-added analysis

  1. adding covariates

  2. combining the approach with interrupted time-series analysis

6. Combining value-added analysis with interrupted time-series analysis

- logic of the combined approach (Figure 5)

- Impact estimate = future deviation from past *pattern* of student gains

[1] From Bloom, Howard S. (1984) “Estimating the Effect of Job-Training Programs Using Longitudinal Data: Ashenfelter’s Findings Reconsidered” ***Journal of Human Resources***, Vol. XIX, No 4, Fall.

- Strengths of the combined approach
    - Value-added analysis helps to control for compositional shifts
    - Interrupted time-series analysis helps to control for selection/maturation bias

- Ways to further strengthen the combined approach
    - Independent replication
    - Comparison series
    - Additional covariates

METHODOLOGICAL UPDATE FROM THE MDRC EVALUATION OF ACCELERATED SCHOOLS 9-13-00

1. Estimating a Counterfactual from Baseline Test Scores

- Using the baseline mean *versus* the baseline trend

- Using only three baseline years

- Not adjusting for “cohort effects”

2. Estimating Program Impacts on the *Distribution* of Test Scores

- Impacts on mean scores

- Impacts on the distribution of scores across “baseline quartiles”

- Impacts on the standard deviation of test scores

Figure 1

A Value-Added Estimate
of Program Impacts
on Student Achievement

Mean Gain for Program Group

$\bar{G}_p$

Mean Gain for Comparison Group

$\bar{G}_c$

$\bar{G}_p = \bar{Y}_{pt} - \bar{Y}_{p(t-k)}$

$\bar{G}_c = \bar{Y}_{ct} - \bar{Y}_{c(t-k)}$

$\hat{I}$ = the estimated impact on value-added

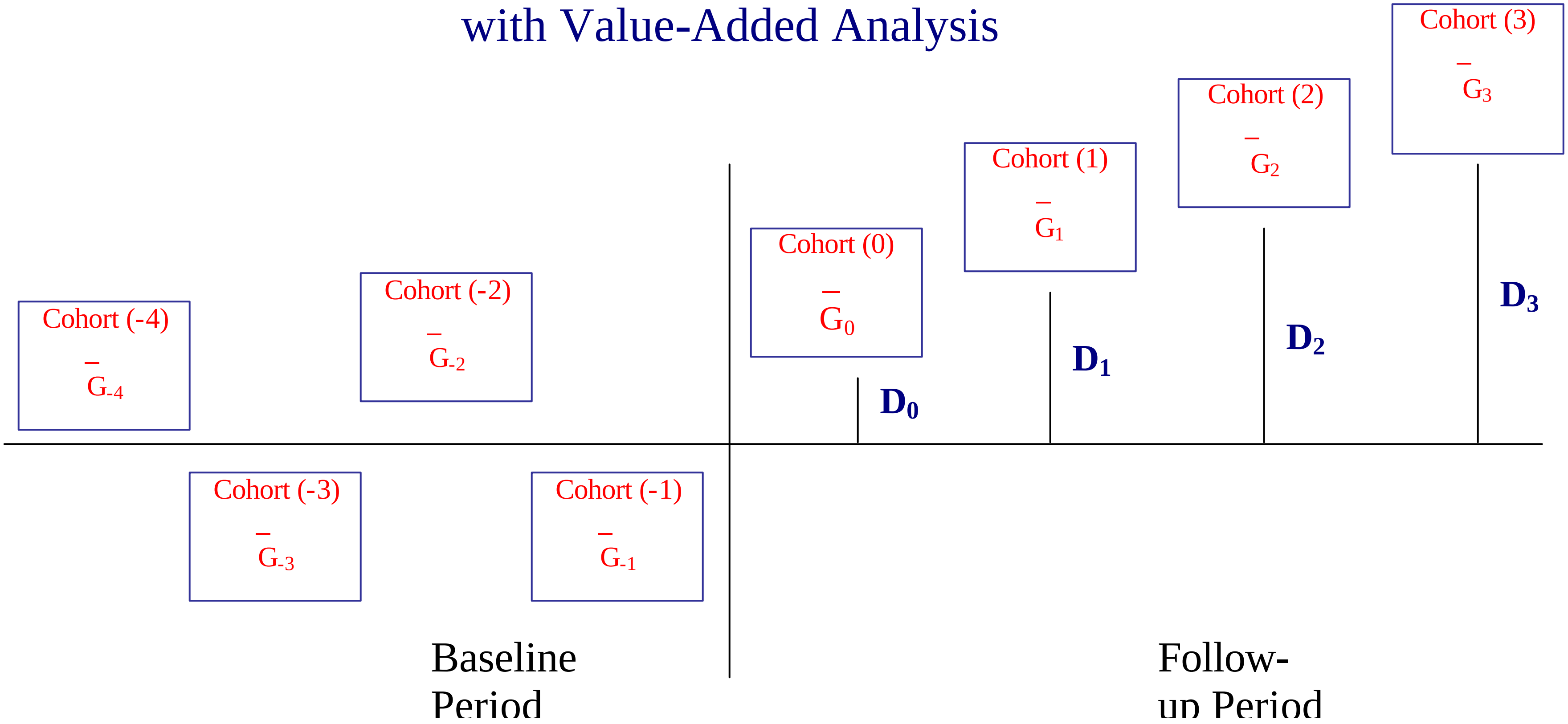
= the difference between mean gains for the program and comparison groups

$= \bar{G}_p - \bar{G}_c$

Figure 2

Combining Interrupted Time-Series Analysis

with Value-Added Analysis



Cohort (-4) $\bar{G}_{-4}$

Cohort (-3) $\bar{G}_{-3}$

Cohort (-2) $\bar{G}_{-2}$

Cohort (-1) $\bar{G}_{-1}$

Cohort (0) $\bar{G}_{0}$

Cohort (1) $\bar{G}_{1}$

Cohort (2) $\bar{G}_{2}$

Cohort (3) $\bar{G}_{3}$

D_0

D_1

D_2

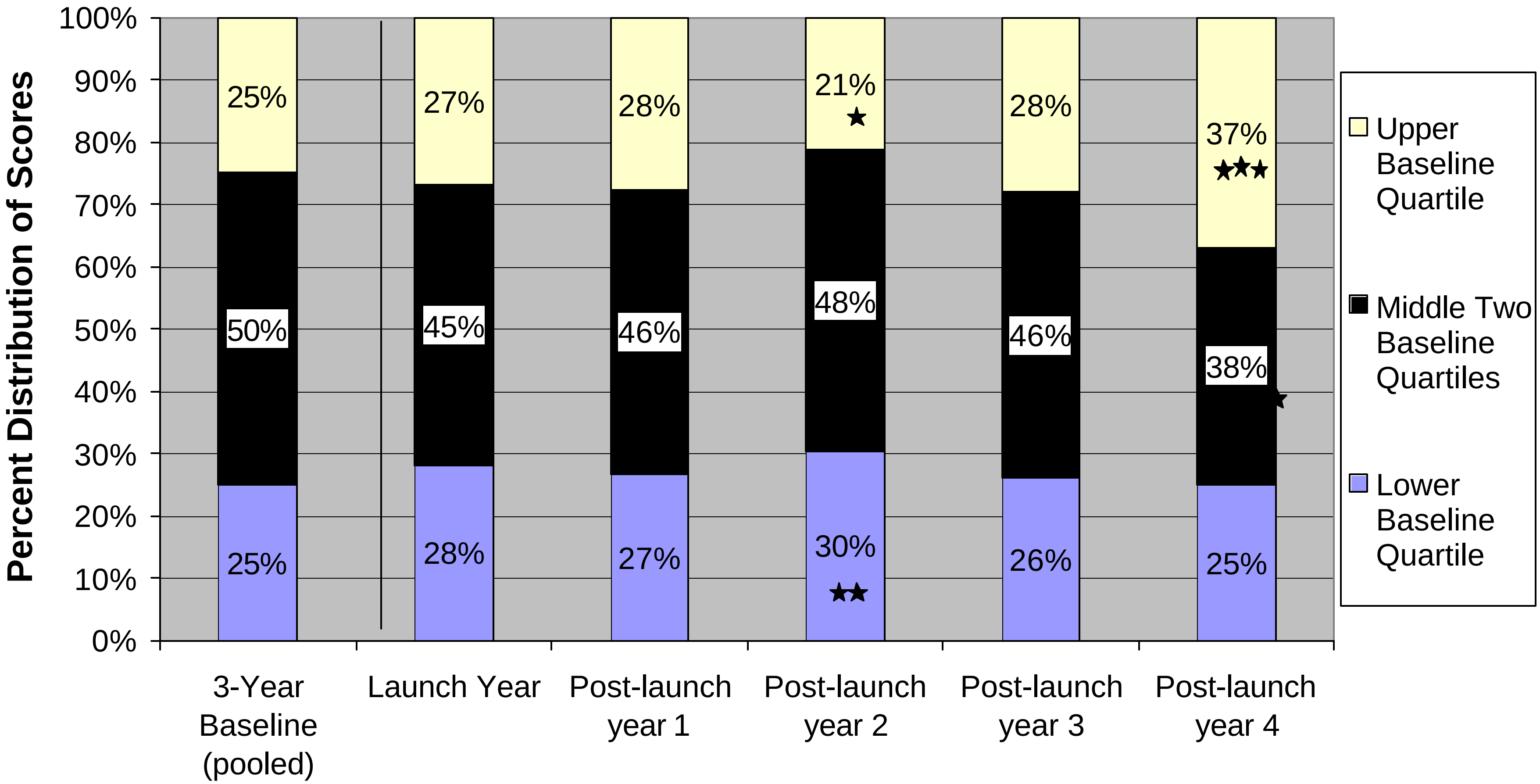
D_3

Baseline Period

Follow-up Period

Impact Estimate = **Future deviation from past pattern of student achievement gain**

A Hypothetical Example of Program Impacts on the Distribution of Student Test Scores



Exerpt on “School Records Research”

from

MDRC

The Evaluation of First Things First:
Research Design Report
March 31, 2000

6.1 General Approach

This section outlines our general approach to estimating the impacts of First Things First. By impact we mean the change in student outcomes *caused* by the initiative, which represents the difference between outcomes experienced in its presence and in its absence (the counterfactual).

The best available strategy for estimating the impacts of First Things First is a combination of the strongest existing quasi-experimental evaluation methods.[2] Thus, we propose to build on the combined strengths of: (1) value-added analysis (Meyer, 1997); (2) hierarchical or multi-level modeling (Bryk and Raudenbush, 1992); and (3) interrupted time-series analysis (Bloom, 1999). Further strengthening our design is the fact that it is based on an explicit theory of change (discussed earlier). Although ideally we would combine *all* of these strategies, we recognize that this might not be possible. Thus, we expect to use a *mix* of strategies that will vary across sites.

In the sections below, we briefly describe each of our planned analytic approaches, noting its key limitations, and indicating how by combining approaches we can address these limitations.

6.1.1 Value-added Analysis. This approach is frequently used to measure student achievement as a function of educational inputs (Meyer, 1997). In its simplest form, value-added analysis represents a post-test/pre-test design with a comparison group. Thus, it *identifies* program impacts on post-test scores by controlling statistically for pre-test scores and student background characteristics.[3]

For example, one might compare eighth grade math achievement scores for program schools with those for comparison schools, controlling statistically for each

[2] We are relying on quasi-experimental methods because it is not feasible to use a randomized experiment to evaluate First Things First.

[3] The most important feature of value-added analysis is controlling for pre-test scores. Controlling for student background characteristics may not add much beyond this.

student's sixth grade scores and background characteristics.[4] Doing so would reflect the difference between program schools and comparison schools in their *increments* to math achievement (value added) between grades six and eight.

This approach has several important limitations. First it does not account for the fact that students are "clustered" by school and that true impacts probably vary by school (they are heterogeneous).[5] Hence, estimated standard errors will understate the uncertainty about program impacts estimates which, in turn, will overstate their statistical significance. This problem can be remedied by using hierarchical modeling to estimate program impacts (discussed below).

A second problem with value-added analysis is selection bias due to program and comparison group differences in the *slopes* of student growth paths. Even if students are at the same level of achievement in 6th grade, they may be on very different growth paths. To the extent that these paths differ initially between program schools and comparison schools, value-added estimates of program impacts would reflect both these initial differences and any subsequent differences caused by First Things First (its impacts). Thus, value-added estimates of program impacts will be biased. As described later, we propose to address this problem by combining value-added analysis with interrupted time-series analysis.

6.1.2 Value-added Analysis with Hierarchical Modeling. To adapt value-added analysis to an hierarchical or multi-level framework, one could, in theory, specify a separate value-added model for each site (program and comparison school pair) and specify how the parameters of these models vary across sites.[6] This represents two-levels of analysis. The first level is the value-added model for each site.[7] The second level is a series of equations which indicate how the parameters of the value-added models vary across sites.[8] The standard errors of program impact estimates obtained in this way

[4] Mathematically, this can be represented as $Y_{it} = a + b_0 P_i + b_1 Y_{i,t-k} + b_2 Xi + e_{it}$, where Y_{it} = the 8th grade math score for student i; $Y_{i,t-k}$ = the math score for student i in an earlier grade (t-k); P_i = 1 if student i is from a program school, and zero otherwise; Xi = a background characteristic for student i, b_0 = the program impact; b_1 = the coefficient for the previous math score, b_2 = the coefficient for the background characteristic, and e_{it} = a random error term.

[5] Both the fact that students are clustered by schools (they are not sampled independently) and the fact that true impacts vary across schools, influence the standard error of impact estimates in ways that require a multi-level or hierarchical analysis.

[6] This approach was used by Sanders and Horn, 1994.

[7] The level-one model is $Y_{itj} = a_j + b_{0j} P_{ij} + b_{1j} Y_{ij,t-k} + b_{2j} X_{ij} + e_{ijt}$, where j represents a specific program and comparison school pair (a "site") and the other symbols are the same as above.

[8] A simple version of the level-two model is:

$$a_j = a_{00} + u_j$$

$$b_{0j} = b_{00} + v_j$$

$$b_{1j} = b_{10} + w_j,$$

where u_j, v_j, and w_j are random variables, b_{00} is the overall mean impact, and b_{0j} is the impact for site j.

account for both the clustering of students by school and the variation in program impacts across schools. Moreover, if there are enough schools, the approach can be adapted to estimate differences in impacts by type of school.

Unfortunately, hierarchical modeling does not reduce selection bias in value-added analysis due to program and comparison group differences in the slopes of student growth paths. To help address this problem, we propose to use interrupted time-series analysis.

6.1.3 Interrupted Time-series Analysis. This approach *identifies* program impacts on student achievement by comparing the performance of current students in a given grade to the trend in achievement of past student cohorts for that grade (Bloom, 1999). MDRC is using this approach for two evaluations of whole-school reforms (Accelerated Schools and Project GRAD) and is planning to use it for a third (Talent Development Schools).

Exhibit 6.1 illustrates the basic approach. It first uses retrospective data to identify a pre-reform baseline trend in test scores for past cohorts of students in a given grade. It then compares the mean score of students in the grade during a follow-up period that begins when the reform is launched to a counterfactual predicted by extending the baseline trend into the follow-up period. The observed *deviation from trend* for each follow-up year provides an estimate of the program impact for that year.

Consider the following hypothetical example, formulated in terms of 8th grade math scores. One might first estimate the trend in these scores for selected First Things First middle schools during four or five years prior to the initiative[9]. One could then compare the expected mean 8th grade math score based on the trend to the actual mean score for a given follow-up year. The deviation from trend for that year would measure the impact of the initiative in that year.[10]

While this approach does not suffer from the selection bias present in value-added analysis, it does have several other limitations: (1) a potential for local events other than First Things First to change student performance (the issue of "local history"), (2) a need to correct standard errors for the clustering of students and the variation of true impacts across sites, and (3) a potential for changes over time in the mix of students (a compositional shift) to cause changes in their observed performance that may complicate the interpretation of impact estimates.

If major changes unrelated to First Things First occur in program schools, such as district level initiatives, new state standards, or changes in administration at the district or

[9] See Bloom (1999) for a discussion of the number of baseline years required.

[10] This estimate could be obtained from the following model $Y_i = a + b_0 P_i + b_1 t_i + e_i$, where Y_i = 8th grade math test for student i; P_i = 1 if this test occurred during the follow-up year and zero otherwise; t_i = a counter for time, which increments by one for each year and b_0 = the deviation from the baseline trend, which is the program impact estimate.

school level, it may be difficult to determine how much of the observed deviation from trend was caused by First Things First and how much was cause by the unrelated changes.

The best way to address this potential problem in the context of our evaluation design is through a careful empirical analysis of the First Things First Theory of Change, described earlier. This theory posits a sequence of changes in intermediate outcomes (i.e. changes in school operating procedures through implementation of the First Things First Theory of Change, followed by increased student engagement and commitment to school, followed by improvements in their school behavior) which in turn lead to measurable increases in student achievement. If increases in student achievement occur in the presence of the hypothesized preceding changes, then it would seem most plausible to attribute the increased achievement (plus all of the preceding changes) to the First Things First initiative. If student achievement increases without evidence of improvements in earlier outcomes, then it would seem more plausible to attribute the improvement to factors other than the initiative. In either even, however, it will be extremely important, to document any and all changes in the local educational system through our implementation analysis.

Another important way to deal with potential local threats to the validity of our estimates of the impacts of First Things First, is through the replication of our analysis in different sites. To the extent that a consistent pattern of increased student performance across First Things First schools is observed, it becomes more plausible that they were caused by the initiative rather than by idiosyncratic local events.

However, even with a theory of change analysis plus replication of this analysis across sites, one must use an hierarchical model to properly estimate the standard errors of program impact estimates (illustrated in a later section).

Hierarchical models do not, however, address problems which can arise from shifts in student composition that may be confounded with First Things First. Such "compositional shifts" can make it difficult to distinguish between changes in achievement due to changes in the mix of students present versus changes in the performance of students who would have been present anyway.

Specifically, if First Things First works as anticipated, it will tend to keep "higher risk" students (who would have dropped out of school early) in school longer. Thus, it will increase the representation of such students (who are likely to perform poorly on standardized tests) among those tested in later grades. This, in turn, will artificially reduce the estimated impact of the initiative. It is not clear how important this problem will be, but the best way to address it is to combine interrupted time-series analysis with value-added analysis in the context of hierarchical models.

6.1.4 Combining Value-added Analysis with Interrupted Time-series Analysis and Hierarchical Modeling. The approach outlined in this section is to identify the impact of First Things First through the deviation from trend in students' *change in achievement* (value added). For example, instead of examining the time path of average 8th grade math scores (as described earlier), one might examine the time path of 8th grade *achievement gains* (e.g. the difference between 6th grade and 8th grade test scores).[11] One could thus measure the impacts of First Things First as the deviation from trend in 6th-to-8th grade achievement gains. To obtain proper standard errors, this analysis could be specified as a two-level hierarchical model. Level one would comprise an interrupted time-series for student gains by school.[12] Level two would specify how the level-one parameters vary across schools.[13]

Combining the preceding approaches can help to offset the limitations of each. Interrupted time series analysis can reduce the potential for selection bias in value-added analysis.[14] Value-added analysis can reduce potential problems due to "compositional shifts" in interrupted time-series analysis.[15] Hierarchical modeling can correct the standard errors for both types of analysis. Furthermore, using multiple sites and theory of change analysis can help to reduce a design's vulnerability to problems of local history.

6.1.5 Statistical Power of the Combined Design. Exhibit 6.2 provides a rough indication of the likely statistical power of First Things First impact estimates from a combined value-added/interrupted time-series analysis. This information is reported in terms of minimum detectable effect size (MDES), which is the smallest true impact that an evaluation design has a "good" chance of detecting. An MDES is expressed as a multiple of the standard deviation of an outcome. Thus, for example, a design with an MDES of 0.25 has a good chance of detecting a true impact that is equal to 0.25 of a standard deviation.

Each MDES in the exhibit assumes impact estimates for a single grade per school with 300 students per grade. Thus, the MDES for one First Things First school, without a

[11] To simplify the discussion, we specify value-added over a given span of grades as the corresponding gain in test scores. In practice, we would use a more flexible approach that specifies the post-test as a dependent variable and the pre-test as a covariate in a regression-type model.

[12] The level one model is thus $\Delta Y_{ij} = a_j + b_{0j} P_{ij} + b_{1j} t_{ij} + b_{2j} X_{ij} + e_{ij}$, where ΔY_{ij} the achievement *change* for student i at school j and P_{ij} = 1 if this change occurred during the follow-up year, t_{ij} = a counter for time, and X_{ij} is a student background characteristic.

[13] One way to represent the level-two model is

$a_j = a_{00} + u_j$

$b_{0j} = b_{00} + v_j$

$b_{1j} = b_{10} + w_j$, where u_j, v_j, and w_j are random variables, b_{00} is the overall mean impact, and b_{0j} is the impact in site j.

[14] While this does not entirely correct for selection, it eliminates the most plausible source of selection bias—selection that is correlated with the gain in student achievement.

[15] It is still possible that these estimates are affected by a shift over time in the *gain* in test scores among student cohorts at First Things First schools, but it is less plausible that this would be large enough to have a meaningful effect on the estimated impacts.

comparison school, reflects 300 students per year. The corresponding MDES for 5 and 10 First Things First schools reflect 1,500 and 3,000 students per year, respectively.[16]

The top panel in the table reports the MDES when pre-test and post-test data can be obtained for *four* annual baseline (that is "pre-reform") cohorts. The second panel reflects corresponding information for *five* annual baseline cohorts. Each column in the table represents a specific follow-up year.

As can be seen, the MDES is larger (the statistical power of impact estimates is lower) for later follow-up years. This reflects the corresponding increase in uncertainty about forecasts from the baseline trend. In addition, the MDES for a four-year baseline trend is larger than its counterpart for a five-year trend. This reflects the greater uncertainty of forecasts from a four-year trend.[17] Lastly, note that the MDES is larger when comparison schools are included than when they are not. This reflects a tradeoff between one's ability to detect differences that might exist and one's ability to infer that First Things First caused these differences.

Although interpretations of effect sizes are somewhat arbitrary, researchers often use Cohen's (1988) rule of thumb that effect sizes of roughly 0.20, 0.50 and 0.80 are small, moderate and large, respectively. Lipsey (1990) provides empirical support for this interpretation based on 102 meta-analyses of studies that mainly involve educational programs.

Given these rough guidelines, estimates in the table suggest that our proposed combined design should have enough statistical power to detect *small average effects* for a total of 10 First Things First high schools or middle schools, if their findings can be pooled across the Kansas City and expansion sites.

Results in the table also suggest that the combination design might have enough power to detect *moderate average effects* for sub-samples of roughly five high schools or middle schools (for example those from only Kansas City or from only the expansion district).

On the other hand, it probably will only be possible to detect quite *large effects* for a single school. This has important implications for how we should view the results for our two new free-standing urban high schools and our two new free-standing rural high schools. In particular, it suggests that being able to pool impact findings, at least within each of these two categories is very important, because we will have very limited statistical power to detect program impacts at a single school.

[16] As discussed later, we do not expect to obtain test score data for the same grade from all First Things First schools. Instead, we hope to be able to pool impact estimates for different grades across schools.
[17] The difference between the statistical power of the two trends is greater for later follow-up years because the uncertainty about these forecasts becomes more important in later years.

Exhibit 6.2

Minimum Detectable Effect Size (MDES)
for Combined Value-added and Interrupted Time-series
Impact Estimates by Grade

	Follow-up Year			
	One	**Two**	**Three**	**Four**
	With a Four Year Baseline Trend			
One First Things First School				
without a comparison school	0.49	0.60	0.72	0.84
with a comparison school	0.70	0.85	1.01	1.19
Five First Things First Schools				
without comparison schools	0.22	0.27	0.32	0.38
with comparison schools	0.31	0.38	0.45	0.53
Ten First Things First Schools				
without comparison schools	0.16	0.19	0.23	0.27
with comparison schools	0.22	0.27	0.32	0.38
	With a Five Year Baseline Trend			
One First Things First School				
without a comparison school	0.45	0.52	0.60	0.68
with a comparison school	0.64	0.73	0.85	0.96
Five First Things First Schools				
without comparison schools	0.20	0.23	0.27	0.30
with comparison schools	0.29	0.33	0.38	0.43
Ten First Things First Schools				
without comparison schools	0.14	0.16	0.19	0.22
with comparison schools	0.20	0.23	0.27	0.31

Assumptions

- 300 students per grade in a school,
- an r-square of 0.45 between post-tests and pre-tests (based on MDRC research using sixth grade and fifth grade test scores for students from 25 Rochester, New York schools in 1991 and 1992),
- a year-to-year intra-class correlation of 0.03 due to differences in annual student cohorts (based on MDRC research using sixth grade and third grade test scores for 25 Rochester schools during the period 1989 - 1992 (Bloom, 1999)),
- average results for five or ten schools *do not* account for the variation in *true impacts* across First Things First schools, which is unknown at this time but would be reflected by estimates using hierarchical modeling,
- one comparison school per First Things First school, and
- minimum detectable effect sizes are reported for a one-tail hypothesis test at the 0.05 significance level with 80 percent statistical power.

Day 4
September 15, 2000

Session 2: Deciding When to Evaluate: MDRC's Report on the Evaluability of the Toyota Family Literacy Program

Goals

We began the workshop series with the idea that evaluation resources are precious and should not be squandered. This session, which focused on a recent MDRC working paper directed toward program operators, illustrated the considerations involved in deciding whether or not to undertake an evaluation of program impacts. The working paper brought together a number of concepts that had been developed during the workshop series, including: the role of program theory in evaluations, the importance of adequate implementation, and requisite sample sizes for studying program impacts.

Topics

I. Selecting program sites to afford a "fair test" of the program model,

II. Using implementation data to investigate the extent of service receipt,

III. Choosing a research design to measure program impacts,

IV. Employing effect size as a common metric across different impact measures,

V. Calculating the requisite sample size given varying assumptions about effect size and other parameters,

VI. Using prior evaluation studies to decide whether an intervention is likely to have effects,

Reading

Janet Quint with Anne Sweeney. (forthcoming) *"An Evaluability Assessment of the Toyota Families in Schools Program"* (New York: MDRC).

CONDUCTING AN EVALUABILITY ASSESSMENT

1. Does the program model make sense?
   What is the underlying theory of change?

2. Is there evidence supporting the model and the theory?

3. What was the implementation experience?
   What quantitative and qualitative data are available to answer that question?

4. What research design is most appropriate for assessing program impacts?

5. What sample sized would be required?

6. Is it likely that the intervention can achieve impacts that are statistically significant and policy-relevant?

DAY 4
Session 3: Random Assignment of Schools To Measure Program Impacts on Student Performance

Goals

This workshop session examined the potential for using random assignment of schools ("cluster assignment") to measure the impacts of educational programs on student performance. The session was designed to introduce workshop participants to the basic concepts of cluster assignment, provide them with intuition about the statistical theory which underlies this approach, indicate how the approach affects the statistical power of program impact estimates, illustrate the statistical power one might expect for a given number of sample schools (based on empirical findings from methodological research conducted by MDRC), and consider what these findings suggest for the number of schools that would be required for a program impact study.

Topics

- Why and when might it be appropriate to randomly assign groups (use cluster assignment) instead of individuals to measure program impacts? In what settings has this been done?
- Why does this approach produce unbiased program impact estimates?
- Why does the approach have less statistical power than random assignment of individuals (for the same number of individuals)?
- What factors affect the statistical power of impact estimates from a cluster assignment design? How might adjustment for baseline covariates (for individual sample members and/or for clusters) reduce the extent to which cluster assignment reduces statistical power?
- Based on data for individual math and reading test scores for third graders and six graders from Rochester, New York in four different years, what is the likely statistical power (minimum detectable effect) of program impact estimates for a study that randomly assigned schools?
- Based on the preceding findings, how many schools would be needed to provide adequate statistical power for an evaluation of the impacts of an educational program on student performance?

Reading

Bloom, Howard S., Johannes M. Bos and Suk-Won Lee (1999) "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs", ***Evaluation Review***, Vol. 23, No. 4, August, pp. 445-469.

Random Assignment of Schools to Measure Program Impacts on Student Performance[18]

Howard S. Bloom
MDRC
June 21, 2000

1. The Generic Evaluation Challenge

- to measure the impacts of programs targeted on whole groups

- examples include evaluations of:
    - o whole school reforms
    - o comprehensive community initiatives
    - o educational technology innovations

[1] Based on Bloom, Howard S., Johannes M. Bos and Suk-Won Lee "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs" (1999) ***Evaluation Review***, Vol. 23, No. 4, August, pp. 445-469.

2. The Basic Approach

- random assignment of groups (cluster assignment)

- comparison of program and control group outcomes

- control for individual-level and/or group-level *covariates* (background characteristics and prior performance)

3. Estimating Impacts Without Covariates

$$\mathbf{Y_{ij}} = a + B_0 \, \mathbf{P_{ij}} + \mathbf{e_j} + e_{ij} \qquad (1)$$

where:

Y_{ij} = the post-test for individual i from school j,
α = the mean post-test for the control population,
B_0 = the true program impact,
P_{ij} = one for students in program schools and zero for students in control schools,
e_j = the error component for school j,
ε_{ij} = the error component for student i from school j.

3. (continued)

$$SE(b_0)_{cluster} = \sqrt{\frac{4t^2}{J} + \frac{4s^2}{nJ}} \qquad (2)$$

where

$SE(b_0)_{cluster}$ = the standard error of the impact estimator using cluster random assignment

J and n = the total number of schools and the number of students per school, respectively

τ^2 and σ^2 = the variance of mean outcomes across schools and across students within schools, respectively

- the intra-class correlation (ρ)

$$r = t^2/(t^2 + s^2) \qquad (3)$$

4. **Estimating Impacts with Covariates**

$$\mathbf{Y_{ij}} = a + B_0\,\mathbf{P_{ij}} + B_1\mathbf{X_{ij}} + \mathbf{e_j} + e_{ij} \qquad (4)$$

- covariates can include individual characteristics and/or school characteristics[19]
- covariates can include demographics, pre-test scores and average test scores of previous cohorts in the same grade from each school
- covariates can differ in terms of how recent they are relative to one's outcome measure
- covariates can reduce τ^2, σ^2 and ρ
- hence, covariates can reduce the minimum detectable effect size and thereby increase the statistical power of a cluster assignment design

[2] Only one covariate, X_{ij}, was included in Equation 4 to simplify the notation.

5. Our Research Questions

- By how much does cluster assignment reduce statistical power without covariates?
- By how much do data on individual pre-tests improve statistical power?
- By how much do data on the mean performance of previous cohorts from each school improve statistical power?
- How does the “recency” of individual or aggregate prior test scores affect their ability to improve statistical power?
- How do answers to the preceding questions vary by *grade, subject and year*?
- **BOTTOM LINE** = How many schools are needed to provide adequate statistical power for a cluster assignment design intended to measure program impacts on student performance?

6. Some of Our Findings (See Table 1)

Table 1
Estimated Minimum Detectable Effect Sizes
For Cohort Approaches
(Table 3 from Bloom, Bos and Lee, 1999) [1]

	Third-Grade		Sixth-Grade		Mean
	Math	**Reading**	**Math**	**Reading**	
Model 1 (no covariates)					
10 schools	0.67	0.65	0.65	0.54	0.63
20 schools	0.47	0.46	0.46	0.38	0.44
30 schools	0.38	0.37	0.38	0.31	0.36
40 schools	0.33	0.32	0.33	0.27	0.31
60 schools	0.27	0.26	0.27	0.22	0.26
Model 2 (Y_{jt-1})					
10 schools	0.37	0.43	0.43	0.33	0.39
20 schools	0.25	0.29	0.29	0.23	0.26
30 schools	0.20	0.23	0.23	0.18	0.21
40 schools	0.17	0.20	0.20	0.16	0.18
60 schools	0.14	0.16	0.16	0.13	0.15
Model 3 (Y_{jt-1}, Y_{jt-2})					
10 schools	0.37	0.32	0.41	0.36	0.37
20 schools	0.24	0.21	0.26	0.23	0.24
30 schools	0.19	0.16	0.21	0.19	0.19
40 schools	0.16	0.14	0.18	0.16	0.16
60 schools	0.13	0.11	0.15	0.13	0.13
Model 4 (Y_{jt-2})					
10 schools	0.43	0.35	0.47	0.38	0.41
20 schools	0.29	0.24	0.32	0.26	0.28
30 schools	0.24	0.19	0.26	0.21	0.22
40 schools	0.20	0.17	0.22	0.18	0.19
60 schools	0.17	0.13	0.18	0.15	0.16
Model 5 (Y_{jt-2}, Y_{jt-3})					
10 schools	0.44	0.33	0.44	0.42	0.41
20 schools	0.28	0.21	0.28	0.27	0.26
30 schools	0.23	0.17	0.22	0.21	0.21
40 schools	0.19	0.15	0.19	0.18	0.18
60 schools	0.16	0.12	0.16	0.15	0.14

[1] The minimum detectable effect size equals the minimum detectable effect measured in raw PEP test scores divided by the standard deviation of the raw scores.

NOTE: Based on the mean values of t^2, s^2 and the sample standard deviation for all years of available full-sample data for each model, and assuming 60 students per school (approximately the average grade size for the full-sample).

7. Further Questions

- How do our findings apply to other standardized tests?
- How do our findings apply when post-tests differ from pre-tests?
- How do our findings apply to studies conducted in more than one city?
- How sensitive is cluster assignment to “contamination of the treatment”?
- How sensitive is cluster assignment to experimental attrition?
- How sensitive is cluster assignment to “outliers”?
- How will our findings from Rochester, NY generalize to other school systems?