

MDRC Working Papers on Research Methodology

**Sample Design for an Evaluation
of the Reading First Program**

**Howard S. Bloom
MDRC**

March 14, 2003

This paper was written for the U.S. Department of Education, Institute of Education Sciences. The author thanks Tracy Rindzius, Jim Kemple, Beth Gamse, and Larry Bernstein for their helpful feedback on an earlier draft plus Corrine Herlihy and Judy Scott for producing the empirical results that are used. All views expressed are those of the author and do not necessarily reflect the official position of the funding organization.

Abstract

This paper illustrates how to design an experimental sample for measuring the effects of educational programs when whole schools are randomized to a program and control group. Questions addressed by the paper include: How many schools should be randomized? How many students per school are needed? What is the best mix of program and control schools? And how do data on aggregate school-level measures of past student performance or individual student-level measures improve the precision of program impact estimates? Empirical analyses based on extensive data from two urban school districts are used to address each question, and the statistical theory underlying these analyses is presented in an appendix. The paper was prepared to help design the evaluation of a national educational program, and it was circulated by the U.S. Department of Education as methodological background for two recent Requests for Evaluation Proposals.

Contents

Abstract	iii
List of Tables and Figures	vii
Introduction	1
Promoting Precision	1
Reporting Precision as Minimum Detectable Effects	2
Precision for Impacts on Average Third Grade Achievement	3
Precision for Impacts on Third Grade Success Rates	11
Precision for Impacts on First-to-Third Grade Achievement Trajectories	14
Precision for Impacts on Instruction	15
Precision and Sample Attrition	17
Promoting Generalizability	19
Promoting a Strong Treatment Contrast	22
Preliminary Recommendations and Next Steps	22
Appendix A: Computing Minimum Detectable Effect Sizes	25
References	33

List of Tables and Figures

Table

1	Minimum Detectable Effect Sizes <i>for Impacts on Third Grade Reading Achievement</i> Given the Presence or Absence of Baseline Achievement Data and the Number of Schools Randomized	5
2	Minimum Detectable Effect Sizes <i>for Impacts on Third Grade Reading Achievement</i> Given the Number of Third Graders per School and the Number of Schools Randomized	6
3	Minimum Detectable Effects and Effect Sizes <i>for Impacts on Third Grade Success Rates</i> Given the Number of Third Graders per School and the Number of Schools Randomized (for District B)	13
4	Minimum Detectable Effect Size <i>for Impacts on the Slopes of First-to-Third Grade Achievement Trajectories</i> Given the Number of Third Graders per School and the Number of Schools Randomized (for District B)	14
5	Minimum Detectable Effect Size <i>for Impacts on Teacher Reports About Local Instruction</i> Given the Number of Teachers Reporting per School and the Number of Schools Randomized (for District C)	16
A.1	Student and School Variance Components for Third Grade Reading Scores (for District A)	27
A.2	Estimated Variance Components for Student Achievement Measures (from District B)	29
A.3	Estimated Variance Components for Teacher Reports on Instruction (from District C)	30

Figure

A.1	Assumed Distribution Across Schools of the True Effect Sizes of Reading First	30
A.2	The Minimum Detectable Effect Multiplier	31

Introduction

The forthcoming evaluation of the federal Reading First program will use a research design that randomly assigns schools in the study sample to a program group, which receives Reading First funds and assistance or a control group, which does not. This approach can provide internally valid estimates of the impacts of Reading First on students by comparing future student outcomes for program schools and control schools. A first necessary step in developing this evaluation is to devise a plan for recruiting, selecting and randomizing schools. To do so one must address three main questions.

- How many schools are needed to detect policy relevant effects on student achievement in grades K–3?
- What is the appropriate ratio of program schools to control schools?
- How many and what types of states and districts are needed to provide an appropriate number of schools plus adequate diversity on key student and school characteristics?

The first two of these questions relate to factors that influence the precision of program impact estimators or, expressed another way, the minimum program impacts that they can detect reliably. The third question relates to how broadly representative these estimators will be, or in other words, their generalizability. The two main sections of this paper therefore address issues related to precision and generalizability for alternative sample designs for the Reading First evaluation. A third section addresses key site features necessary to provide a fair test of the program. The final section summarizes our preliminary recommendations for an evaluation sample and outlines our next steps for developing a site recruitment plan.

Before proceeding however, it should be noted that final sample decisions must also reflect the research design and analysis plan chosen for the Reading First evaluation. These additional decisions must address issues such as the types of measures to be taken, when and how they will be taken, for whom they will be taken and how they will be analyzed. Thus, the present paper provides one important source of input to a more comprehensive set of evaluation decisions.

Promoting Precision

Our discussion of precision proceeds as follows. We first describe how to measure and assess the precision of program impact estimators in terms of their minimum detectable effects. We then use this metric to explore how sample design can affect the precision of estimates of Reading First impacts on third grade student performance (We focus first on this outcome because of the program's emphasis on it.) We then examine how sample design can affect the precision of impact estimators for other student and teacher outcomes. We lastly consider how sample attrition can affect precision.

Reporting Precision as Minimum Detectable Effects

Minimum detectable effects are a simple way to express the statistical precision of an impact study design. Intuitively, a minimum detectable effect is the smallest program impact that could be measured with confidence given random sampling and estimation error.¹ For example, from a benefit-cost perspective one might ask whether a proposed sample could reliably detect the smallest impact needed for a program to “break even” (that is, produce benefits equal to costs.) One would want a sample that was large enough to ensure that an estimated impact around this “break-even” point was a reliable indicator of the program’s true impact and not just due to chance variation. A smaller sample might only enable the study to detect impacts that are well above this break-even point and thus very difficult to attain. The study would therefore miss an opportunity to produce reliable estimates of policy-relevant impacts. Hence, it would be “under-powered” statistically.

Four main features of the Reading First sample will determine its minimum detectable effects:

- The number of schools randomized to Reading First and control status,
- The availability of baseline data on student achievement,
- The number of students per grade at each school, and
- The mix of Reading First schools versus control schools.

Appendix A explains how these factors are related to the precision of program impact estimators and describes how we used these relationships to assess the likely precision of alternative sample designs. These assessments were based on empirical findings obtained from student test score data for two urban school districts and teacher survey data for a third urban district.²

Before discussing these findings and their implications it is important to understand how we report the precision of program impact estimators. In all but one case, we report precision as “minimum detectable effect sizes.” This metric, which is used widely for measuring the impacts of educational programs, is defined in terms of the underlying population standard deviation of student achievement. For example, a minimum detectable effect size of 0.25 indicates that an impact estimator can reliably detect a program-induced increase in student achievement that is equal to or greater than 0.25 standard deviations of the existing student distribution.

Although judgments about whether a specific effect size is large or small are somewhat arbitrary, several useful guidelines exist. Many researchers use a rule of thumb proposed by Cohen (1988) who suggested that effect sizes of roughly 0.20 be considered small, 0.50 be considered moderate, and 0.80 be considered large.³ Lipsey (1990)

¹We define a minimum detectable effect as the smallest true program impact that would have an 80 percent chance of being detected (have 80 percent power) using a one-tail hypothesis test at the 0.05 level of statistical significance. We use a one-tail test because the central policy issue to be addressed by the Reading First evaluation is whether the program *improves* student reading performance.

²We do not identify each district by name.

³Cohen (1988).

provides empirical support for this typology based on the distribution of 102 mean effect sizes obtained from 186 meta-analyses of treatment effectiveness studies, most of which come from education research.⁴ The bottom third of his observed distribution (small effects) ranges from zero to 0.32; the middle third (moderate effects) ranges from 0.33 to 0.55; and the top third (large effects) ranges from 0.56 to 1.26.

Another benchmark for assessing effect size magnitudes is the Tennessee elementary school class size experiment, Project STAR. This study is relevant to the Reading First evaluation for three reasons. First, it represents a major effort to improve elementary school student performance. Second, it estimates program impacts using a random assignment research design. Third, it defines effect size in the same way that it probably will be defined for Reading First. Program impact estimates for Project STAR range mainly from about 0.15 to 0.25 standard deviations.⁵ This observed impact on average student reading and math achievement was produced by a class size reduction from 22-26 students per normal size class to 13-17 students per reduced size class—a substantial difference.

Precision for Impacts on Average Third Grade Achievement

Table 1 presents our assessment of how minimum detectable effect size for estimates of impacts on average third grade reading achievement would vary with the number of schools in the Reading First evaluation sample. This assessment is based on three different scenarios about the existence of baseline student achievement data. Each scenario assumes a 50/50 mix of Reading First and control schools with 60 third grade students per school.

Findings in the top panel of Table 1 are based on the reading achievement of third grade students from 25 elementary schools in one urban school district.⁶ Findings in the bottom panel are based on the reading achievement of third grade students from 131 elementary schools in another urban school district.⁷ Replicating these findings using data from two different school districts—one from the Northeast and one from the Southwest—helps to assess their robustness and enhance their generalizability. This, in turn, should increase one's confidence in using these findings as a basis for designing a Reading First evaluation sample. In future work we hope to further replicate these analyses using data from many different school districts. This research could make an important contribution to expanding the knowledge base for designing other educational evaluations that use random assignment of schools.

The first column in Table 1 lists minimum detectable effect sizes for estimators that do not control for past student reading achievement. The second column lists minimum detectable effect sizes for estimators that control for each school's mean reading test score for the same grade two years before Reading First was launched. Thus

⁴Lipsey (1990).

⁵Finn et al. (1999) and Nye et al. (1999).

⁶These findings were based on results from Bloom et al. (1999).

⁷These findings were produced by MDRC for the present paper.

for example, they might apply to estimates of program impacts on mean third grade reading scores for the second year after Reading First is launched, controlling for schools' mean third grade reading scores in the year before Reading First was launched.

The third column in the table lists minimum detectable effect sizes for estimators that control for individual student-level reading achievement two grades prior. For example, they might apply to estimates of program impacts on mean third grade test scores for the second year after Reading First is launched, controlling for each student's first grade scores (on tests taken the year before Reading First was launched).

Table 2 indicates how minimum detectable effect size varies with the number of schools in the sample and the number of third grade students per school. These results assume a 50/50 mix of Reading First and control schools. In addition, they assume that past student achievement is controlled for by a school-level baseline measure that is lagged two years (i.e. mean baseline third grade reading scores from two years before followup reading scores are measured). Once again, separate results are reported for the two urban school districts for which data were available.

The findings in Tables 1 and 2 plus an examination of the underlying statistical properties of school randomization suggest several important lessons for the Reading First evaluation sample.

Lesson #1: Aggregate school-level data or individual student-level data on baseline reading achievement can markedly improve precision. Table 1 indicates that such baseline data can reduce minimum detectable effect sizes by roughly 30 percent (for District B) to 40 percent (for District A). This is equivalent to increasing the number of schools randomized by roughly 2 to 2.5 times, respectively. And this result holds regardless of the number of schools randomized. For example, according to findings for District A it would take 50 randomized schools to attain a minimum detectable effect size of 0.32 without baseline student achievement data; but with such data it would only take 20 schools. Likewise, according to findings for District B, it would take 60 randomized schools to attain a minimum detectable effect size of 0.32 without past student achievement data but only 30 randomized schools with such data.

Table 1

**Minimum Detectable Effect Sizes
for Impacts on Third Grade Reading Achievement
Given the Presence or Absence of Baseline Achievement Data
and the Number of Schools Randomized**

Schools Randomized	No Baseline Achievement Data	School-level Baseline Achievement Data Lagged Two Years	Individual-level Baseline Achievement Data Lagged Two Years
Urban District A			
10	0.79	0.48	0.50
20	0.53	0.32	0.33
30	0.42	0.26	0.27
40	0.36	0.22	0.23
50	0.32	0.20	0.21
60	0.29	0.18	0.19
70	0.27	0.17	0.17
80	0.25	0.16	0.16
Urban District B			
10	0.86	0.61	0.66
20	0.57	0.41	0.44
30	0.46	0.33	0.36
40	0.40	0.28	0.31
50	0.35	0.25	0.27
60	0.32	0.23	0.25
70	0.30	0.21	0.23
80	0.28	0.20	0.21

NOTE: Minimum detectable effect sizes are for one-tail hypothesis tests at the 0.05-significance level with 80 percent power for a 50/50 mix of Reading First and control schools with 60 third grade students per school.

Table 2

**Minimum Detectable Effect Sizes
for Impacts on Third Grade Reading Achievement
Given the Number of Third Graders per School
and the Number of Schools Randomized**

Schools Randomized	Third graders per school					
	20	40	60	80	100	120
Urban District A						
10	0.56	0.50	0.48	0.47	0.46	0.46
20	0.38	0.34	0.32	0.32	0.31	0.31
30	0.30	0.27	0.26	0.25	0.25	0.25
40	0.26	0.23	0.22	0.22	0.21	0.21
50	0.23	0.21	0.20	0.19	0.19	0.19
60	0.21	0.19	0.18	0.18	0.17	0.17
70	0.20	0.17	0.17	0.16	0.16	0.16
80	0.18	0.16	0.16	0.15	0.15	0.15
Urban District B						
10	0.67	0.63	0.61	0.60	0.60	0.59
20	0.45	0.42	0.41	0.40	0.40	0.40
30	0.36	0.34	0.33	0.32	0.32	0.32
40	0.31	0.29	0.28	0.28	0.28	0.27
50	0.28	0.26	0.25	0.25	0.25	0.24
60	0.25	0.23	0.23	0.23	0.22	0.22
70	0.23	0.22	0.21	0.21	0.21	0.21
80	0.22	0.20	0.20	0.19	0.19	0.19

NOTE: Minimum detectable effect sizes are for one-tail hypothesis tests at the 0.05-significance level with 80 percent power for a 50/50 mix of Reading First and control schools with school-level baseline achievement data lagged two years.

Another important implication of the findings in Table 1 is that aggregate school-level baseline achievement data can improve precision by as much as can individual student-level data. (Compare columns two and three in both panels.) Thus, it might be possible to use existing summaries of past student test scores for some or all of the districts in the Reading First evaluation to attain the precision required.⁸ This might be especially important in light of the fact that few states require testing in grades K or one (which is necessary for a student-level baseline covariate) but virtually all states require testing in grade three (which can provide a school-level baseline covariate).

One final point that is not illustrated by the table but was demonstrated by Bloom et al. (1999) using data for District A is that precision gains from baseline achievement data do not necessarily depend on whether the same test was used to measure baseline and followup outcomes.⁹ This provides yet another degree of flexibility for constructing the Reading First sample.

Lesson #2: Randomizing 50 to 70 schools should provide adequate precision for overall program impact estimates given baseline achievement data.

Reading down any column in Table 1 indicates that increasing the number of randomized schools greatly reduces minimum detectable effect size until one reaches about 50 to 70 schools. After this point, there is a pronounced diminishing marginal return to additional schools. Thus if data collection costs are roughly proportional to the number of schools in the sample (because they do not benefit appreciably from economies of scale), it may not be cost-effective to go much beyond 50 to 70 schools for the Reading First evaluation—at least for estimating overall program impacts.

On the other hand, if the marginal costs of adding schools are fairly small, including more than 70 schools might provide a feasible way to improve subgroup impacts estimates, which will be less precise than those for the overall sample (discussed later). In addition, choosing a somewhat larger sample of schools can help to buffer the evaluation from inevitable complications that will reduce its precision, such as sample attrition, missing data, or dilution of the treatment contrast (also discussed later).

Reading down column two in Table 1 also indicates that with 50 to 70 schools, minimum detectable effect sizes range from about 0.20 to 0.17 for District A or 0.25 to 0.21 for District B given baseline achievement data. These minimum detectable effects are comparable to the actual impacts produced by the Tennessee

⁸If schools are randomized separately within each school district it is possible to accommodate different baseline tests for different districts. Furthermore, it is possible to accommodate school-level baseline data for some districts and student-level baseline data for others. These decisions must be addressed as part of the research design for the Reading First evaluation.

⁹Bloom et al. (1999) find that precision for estimating impacts on third grade reading achievement is roughly the same regardless of whether one controls for baseline *reading or math* scores. This provides strong evidence that precision for program impact estimators can be improved markedly even when the baseline reading test differs from the followup reading test.

class-size experiment. Hence, they are likely to be policy-relevant. Furthermore, they are in range of small impacts according to Cohen (1988) and Lipsey (1990). All three of these benchmarks therefore suggest adequate precision for estimating overall Reading First impacts with a sample of 50 to 70 randomized schools.

Lesson #3: Randomizing 50 to 70 schools should provide adequate precision to estimate impacts for some subgroups of schools given baseline achievement data. This lesson flows from the findings in Table 1 that minimum detectable effect sizes for 30 randomized schools are 0.26 and 0.33 for Districts A and B, respectively given school-level baseline achievement data. Thus, if 50 to 70 schools are randomized, subgroups of roughly half these schools could detect true program impacts that are slightly larger than those for the Tennessee class size experiment. Such subgroups might be defined in terms of school size, racial/ethnic composition, geographic location or other school characteristics.

It is important to note, however, that more stringent limitations will exist on the ability to determine whether impact estimates for subgroups of schools are statistically significantly different from each other. This distinction is doubly difficult to make because there is less precision for estimating the difference in impacts between two groups than there is for estimating the impact for each,¹⁰ and the difference between two impacts is likely to be smaller than each impact. Hence, it will be important to identify in advance the highest priorities for types of school subgroups to focus on in the Reading First evaluation.

A further important point to note is that only very large impacts will be detectable for very small subgroups of schools. For example, Table 1 indicates that only impacts equal to or greater than 0.5 to 0.6 standard deviations (which are quite large and not often found by rigorous evaluations of education programs) could be detected reliably for a subgroup of 10 schools with a 50/50 program/control mix. So there are important limits to the types of school subgroups for which impacts will be estimable.

Lesson #4: School size probably will have almost no influence on the precision of overall Reading First impact estimates. Table 2 demonstrates this point. For example, according to data from District A, the minimum detectable effect size for 60 randomized schools would vary from 0.21 to 0.17 as the number of third graders per school varies from 20 to 120—a six-fold increase.¹¹ The corresponding range for District B is 0.25 to 0.22.¹² Therefore it will not be necessary to exclude small schools from the Reading First evaluation. Conversely,

¹⁰For two subgroups of equal size and composition the minimum detectable difference in subgroup effect sizes equals $\sqrt{2}$ times the minimum detectable effect size for each subgroup.

¹¹Our estimates of minimum detectable effect size follow the approach of Raudenbush (1997) and assume for simplicity a constant number of students per grade at each school. These estimates provide a good planning guide if interpreted as average grade size.

¹²The exact ratios of minimum detectable effect size across columns in Table 2 are the same for all rows although the ratios listed vary somewhat due to rounding.

it will not be necessary to oversample large schools. Indeed, given the likely additional costs of data collection for large schools, it might be inefficient to oversample them.

Lesson #5: Impact estimates for subgroups of students that are well represented at all sample schools may have almost as much precision as those for the full study sample. This point is a corollary of the fact that school size barely affects precision. Consider the case of student subgroups by gender. First note that boys and girls probably will be roughly equally represented at all schools in the sample. Next note that for 60 randomized schools with 60 third graders per school (boys plus girls) the minimum detectable effect size is 0.18 or 0.23 (for Districts A or B, respectively).

So how much precision will there be for 30 third grade boys or 30 third grade girls per school? Interpolating findings in Table 2 for 20 third graders per school and for 40 third graders per school indicates that the minimum detectable effect size for 30 third graders per school is 0.20 or 0.24. Thus, impact estimates for boys or girls separately will be almost as precise as those for boys and girls together.

This result applies to any student subgroup that is well represented at all sample schools (such as students receiving free or reduced-price meals) because the number of schools is what mainly determines precision. Furthermore, statistical significance tests for estimates of differences in program impacts between mutually exclusive student subgroups represented at all sample schools can sometimes have *more statistical power* than do significance tests for full-sample average impact estimates.¹³

It is important to note that these properties do not hold for subgroups that are not well represented at all sample schools, which probably will be the case for racial and ethnic subgroup groups, among others. Impact estimates for these subgroups will involve both fewer schools and fewer students per school. Hence, their precision will be limited accordingly.

Furthermore, and even more problematic for these subgroups, is that differences between their impact estimates might reflect differences in their distributions across schools. Hence unobserved school differences may be confounded with student differences in ways that cannot be separated.

Lesson #6: A 50/50 balanced mix of Reading First and control schools provides the most precise and most robust impact estimators possible. Textbooks often note that a balanced 50/50 sample mix minimizes the standard error of impact estimators and thereby maximizes precision. However unbalanced samples do not increase standard errors appreciably until they begin to exceed a

¹³Bloom (in process) demonstrates this possibility.

roughly 75/25 or 25/75 mix.¹⁴ Hence in this regard, there might be more room for unbalanced designs than often is thought to be the case.

Nonetheless, of greater importance for the Reading First evaluation is that balanced designs are more robust to violations of distributional assumptions that underlie statistical significance tests for impact estimates.¹⁵ Hence, balanced designs are more likely to provide valid tests when these assumptions are violated.

Furthermore, balanced designs make it possible to use the simplest available impact estimators. This avoids *compounding complexity* with analytic issues due to sample attrition, program/control group cross-contamination, missing data and other problems that arise when evaluation designs are implemented in the field. Because for random assignment of schools even the simplest impact estimators require somewhat complex statistical tests, compound complexity is especially problematic and should be avoided in all possible ways.

Fortunately, it is possible to use a balanced sample design for the Reading First evaluation, while at the same time minimizing the number of schools to be randomized to control status (which can increase the willingness of districts to participate in the study).

For this purpose schools that are judged eligible for the Reading First evaluation in a district could be randomized to three groups: (1) a Reading First program group, which would receive program resources and be part of the evaluation, (2) a Reading First control group, which would not receive program resources for a specified period of time but would be part of the evaluation, plus (3) a non-research group that would receive Reading First resources but would not be part of the evaluation. For example, in a district with nine schools judged eligible for the evaluation, three schools could be randomized to each group. Thus, only one third of the eligible schools would be randomized to control status even though the number of schools in the program and control groups would be the same. (Follow-up data collection costs would only be incurred for schools in the program and control groups because only they would be part of the research sample.)

¹⁴Bloom (1995).

¹⁵Gail et al. (1996) demonstrate that if individual error variances are the same for a program and control group then both balanced and unbalanced designs will produce valid significance tests for impact estimates. But if these error variances are different—which will occur if Reading First impacts vary across schools, as is likely (see Appendix A)—then only balanced designs will produce valid test statistics for simple impact estimators. Although more complex analyses can be used for unbalanced designs, they reduce and make it difficult to determine the number of degrees of freedom involved. Furthermore, when they are combined with other analytic problems that arise they make it very difficult, if not impossible to produce valid statistical tests.

Lesson #7: Regardless of initial assumptions or calculations, there is never enough precision to meet all evaluation objectives. From our collective personal and organizational experiences, it seems that precision is always at a premium and there is never enough to go around. Therefore it is important to develop a conservative Reading First sample design that will withstand the challenges of student and school attrition,¹⁶ missing data, dilution of the treatment contrast, and other unforeseen imperfections that inevitably materialize when an evaluation is implemented in the field.

Such a conservative sample design would have three main features:

1. *It would maximize the number of schools included given realistic budget constraints for the evaluation.* Even though our assessments of precision suggest that 50 to 70 schools would be adequate for estimating overall program impacts, it would be prudent to recruit a sample that was on the high side of this range—if not slightly higher. Doing so would help to ensure that the final analysis sample has adequate precision for estimating both overall program impacts and school subgroup impacts.
2. *It would include only schools for which baseline achievement data are available.* As illustrated, such baseline information can markedly improve precision and is necessary in order for the recommended sample size of 50 to 70 schools to be adequate. Without such information a considerably larger sample of schools would be needed to do the same job.
3. *It would randomize a balanced 50/50 mix of program and control schools.* As noted, this sample feature is important to ensure robustness of the impact analysis to a wide range of potential problems and to maintain its simplicity—and thus transparency. The easiest way to implement this feature and at the same time minimize the fraction of eligible schools randomized to control status is to include a non-research group in the randomization. This group could receive Reading First resources but would not be part of the evaluation.

Precision for Impacts on Third Grade Success Rates

Because a central goal of the Reading First program is to have all children reading at grade level by third grade, we assessed the likely precision of Reading First impact estimates for this type of yes/no binary outcome. To do so we examined precision for estimates of impacts on student success rates measured as the percentage of third graders

¹⁶This causes problems for the precision and validity of program impact estimators.

who attained a reading achievement threshold. Appendix A describes how this was done using test score data for District B.¹⁷

Table 3 presents our findings in two different metrics to provide alternative ways of assessing their implications. The top panel of the table reports minimum detectable effects in percentage points. For example, a minimum detectable effect of 11 percentage points means that a sample design can reliably detect a program-induced increase of 11 percentage points or more in student success (such as an increase from 30 to 41 percent). The bottom panel of the table reports minimum detectable effect sizes based on the underlying standard deviation of the binary distribution of student successes. Thus as can be seen, an 11 percentage-point minimum detectable effect implies a 0.22 standard deviation minimum detectable effect size.

Comparing the bottom panel of Table 3 for student success rates with the bottom panel of Table 2 for average student achievement levels indicates that they are almost identical. Hence, all of the lessons discussed earlier also apply to this second way of measuring student achievement.

Furthermore, results in the top panel of Table 3 provide an alternative way to think about the policy implications of precision for potential Reading First sample designs. As can be seen, randomizing 60 schools with a 50/50 program/control mix, 60 third graders per school and school-level baseline achievement data, yields a minimum detectable effect of 11 percentage points or 0.22 standard deviations. The 0.22 effect size is comparable in magnitude to the actual effects produced by the Tennessee class size experiment. In addition it is in the range characterized as small impacts by Cohen (1988) and Lipsey (1990).

But what does it mean to increase the percentage of third grade students who read at grade level by 11 percentage points? Is this a large, moderate or small increase and thus is the degree of precision it represents low, moderate or high? Assessing this magnitude is a policy decision, which should reflect the stated goals of the Reading First program, the likely costs of the program and the ability of past similar programs to improve student reading. Thus, further consideration of this finding by Department staff and officials is important before final sample design decisions are made for the Reading First evaluation.

¹⁷Comparable data were not readily available for District A.

Table 3

**Minimum Detectable Effects and Effect Sizes
for Impacts on Third Grade Success Rates
Given the Number of Third Graders per School
and the Number of Schools Randomized
(for District B)**

Schools Randomized	Third graders per school					
	20	40	60	80	100	120
In Percentage Points						
10	33%	30%	29%	29%	28%	28%
20	22	20	20	19	19	19
30	18	16	16	15	15	15
40	15	14	14	13	13	13
50	14	12	12	12	12	12
60	12	11	11	11	11	11
70	11	10	10	10	10	10
80	11	10	9	9	9	9
In Effect Size						
10	0.65	0.60	0.58	0.57	0.56	0.56
20	0.43	0.40	0.39	0.38	0.38	0.37
30	0.35	0.32	0.31	0.31	0.30	0.30
40	0.30	0.28	0.27	0.26	0.26	0.26
50	0.27	0.25	0.24	0.23	0.23	0.23
60	0.24	0.22	0.22	0.21	0.21	0.21
70	0.23	0.21	0.20	0.20	0.20	0.19
80	0.21	0.19	0.19	0.18	0.18	0.18

NOTE: Minimum detectable effect sizes are for one-tail hypothesis tests at the 0.05-significance level with 80 percent power for a 50/50 mix of Reading First and control schools with school-level baseline achievement data lagged two years and a binary outcome variable.

Precision for Impacts on First-to-Third Grade Achievement Trajectories

Because educational researchers have considerable interest in how programs affect the trajectories of student achievement over time, we also examined likely minimum detectable effects for this type of outcome measure. Appendix A describes how this was done for the *slopes* of individual achievement trajectories using annual first grade to third grade longitudinal reading test score data from District B.¹⁸ Table 4 reports these findings as minimum detectable effect sizes based on the standard deviation of observed student slopes. Findings in the table do not use a baseline covariate because no plausibly effective covariate was readily available nor is one likely to exist for the Reading First evaluation (although further research on this question is warranted).

Table 4

**Minimum Detectable Effect Size
for Impacts on the Slopes of First-to-Third Grade Achievement Trajectories
Given the Number of Third Graders per School
and the Number of Schools Randomized
(for District B)**

Schools Randomized	Third Graders Per School					
	20	40	60	80	100	120
10	0.73	0.68	0.67	0.66	0.65	0.65
20	0.49	0.46	0.45	0.44	0.44	0.43
30	0.39	0.37	0.36	0.35	0.35	0.35
40	0.34	0.32	0.31	0.30	0.30	0.30
50	0.30	0.28	0.27	0.27	0.27	0.27
60	0.27	0.26	0.25	0.25	0.24	0.24
70	0.25	0.24	0.23	0.23	0.23	0.23
80	0.24	0.22	0.22	0.21	0.21	0.21

NOTE: Minimum detectable effect sizes are for a one-tail significance test with a 50/50 mix of Reading First and control schools and no covariate.

So what do these findings suggest? To answer this question compare the findings for slopes in Table 4 with the corresponding findings for success rates in the bottom panel of Table 3 with the corresponding findings for achievement levels in the bottom panel of Table 2. Even though these three outcome measures have different meanings and could have different statistical properties, their estimated minimum detectable effect sizes tell a strikingly similar story. Although further research into why these findings are so similar would be quite helpful and further replication of the findings using data for

¹⁸Once again, comparable data were not readily available for District A.

additional school districts would be important for future evaluations, their bottom lines for the present paper support the seven basic lessons outlined above.

Precision for Impacts on Instruction

In addition to estimating the impacts of Reading First on student achievement it also will be important to examine how the program was implemented and why its implementation varied across schools and districts. Doing so is necessary to understand why Reading First was or was not effective overall and what caused its variation in effectiveness. A central concern in this regard is how the program affected instructional practices—the principal engine of student achievement.

One common source of information about instructional practices is survey responses by teachers to questions about what they and their colleagues do. Another potentially more accurate but far more expensive approach is direct classroom observation. Both types of measures are difficult to construct and limited in important ways. For example, at a conceptual level there are many dimensions to instructional practice and it is extremely difficult to distill them down to one or a few meaningful scales. At an operational level it is difficult to formulate survey questions that elicit accurate responses about specific dimensions. And it is difficult to develop observational protocols that are both accurate and cost-effective. In addition, for observational protocols it is extremely costly to sample enough classrooms enough times for long enough periods to produce reliable results. Thus measuring instructional practices is difficult and costly. Nevertheless, because of instruction's central role in the educational process it is important to find ways to measure program impacts on it.

With this in mind, we estimated minimum detectable effect sizes for composite scales measuring local instruction based on teacher survey responses for seven schools at a third urban school district (District C). Appendix A describes how this was done for three different scales and how their average results were used to compute minimum detectable effect sizes. Table 5 summarizes these findings for different numbers of teachers reporting per school. This makes it possible to examine the likely precision if one limits responses to teachers from a single grade at a time (on the premise that instructional practices differ too much across grades to produce meaningful averages) or if one is willing to pool teacher responses across several grades (on the premise that there are meaningful commonalities). Note that the minimum detectable effects in the table control for school-level average baseline teacher responses lagged one year.

First note that findings in Table 5 for teacher survey responses differ in one important respect from those in Tables 1 – 4 for student achievement—the number of teachers reporting from each school has a pronounced affect on precision, whereas the number of students tested does not.

The reason for this difference in findings is the extreme difference in the range of teachers and students per school. For the very small numbers of teachers employed by most elementary schools (assumed in the table to range from about 3 to 18) precision increases rapidly with an increase in the number of survey responses averaged. But for

the much larger numbers of students who attend most elementary schools (assumed in the table to range from about 20 to 120) precision is less sensitive to how many there are the school average test score. In other words, precision increases markedly as the number of individuals per school increases from about 3 to 20, but does not increase by nearly as much thereafter.

Table 5
Minimum Detectable Effect Size
for Impacts on Teacher Reports About Local Instruction
Given the Number of Teachers Reporting per School
and the Number of Schools Randomized
(for District C)

Schools Randomized	Teachers Reporting Per School					
	3	6	9	12	15	18
10	1.05	0.83	0.75	0.70	0.67	0.65
20	0.70	0.56	0.50	0.47	0.45	0.44
30	0.56	0.45	0.40	0.38	0.36	0.35
40	0.49	0.39	0.35	0.32	0.31	0.30
50	0.43	0.34	0.31	0.29	0.28	0.27
60	0.39	0.31	0.28	0.26	0.25	0.24
70	0.36	0.29	0.26	0.24	0.23	0.23
80	0.34	0.27	0.24	0.23	0.22	0.21

NOTE: Minimum detectable effect sizes are for a one-tail significance test with a 50/50 mix of Reading First and control schools and mean school-level teacher reports for the prior year as a covariate.

With this in mind, note that for 60 randomized schools with 3 teachers per grade, the minimum detectable effect size for a single grade is roughly 0.39 standard deviations. This is a moderate size impact indicating moderate precision according to the benchmarks noted earlier. Next note that if one is willing to pool teacher responses across first, second and third grades (providing nine teachers per school), the minimum detectable effect size drops to 0.28.

When interpreting these findings, several points should be kept in mind. First is that Reading First impacts on teacher outcomes such as instructional practices may well be larger than its impacts on student achievement. This is because the link between program resources and teacher behavior (created by teacher training, coaching, and other forms of professional development) is more direct than is the link between program resources and student achievement. Thus, somewhat larger minimum detectable effects for teacher outcomes might not be problematic. On the other hand, given the difficulties noted above with respect to defining, constructing and administering measures of

instructional practices, considerable work remains before an accurate and cost-effective method for doing so is available.

Precision and Sample Attrition

This final section on precision briefly examines the potential effects of school and student attrition from the Reading First evaluation sample. Four situations are considered: (1) school attrition from the program but not from the evaluation, (2) school attrition from the program and the evaluation, (3) student attrition due to mobility that is independent of the program, and (4) student attrition due to mobility that is affected by the program.

To consider these issues in depth requires thoroughly examining the research design and analysis plan for the Reading First evaluation, which does not exist at this time and is beyond the scope of the present paper. Nevertheless, because these issues are so important, we identify the types of methodological problems they can create in general and discuss their effects on precision in particular.

Case #1: If some schools that were randomized to Reading First drop out of the program but cooperate with data collection, one could produce unbiased impact estimates for the full sample. However, any school that drops out this way will reduce the treatment contrast between program and control schools. This in turn, will dilute average impacts, thereby reducing statistical power. In addition it will complicate interpretation of impact findings because they will represent a mix of schools that did and did not participate in the program.

Bloom (1984) presents a simple adjustment for estimating program impacts on subjects that participate in a program to which they are randomized when the program does not appreciably affect subjects that do not participate. This might apply if some schools are randomized to Reading First but never participate. Hence, the method might produce internally valid estimates of Reading First impacts on schools that participate. Nevertheless, the precision lost by such attrition cannot be recouped. Hence, the minimum detectable effect size for schools that participate equals the full-sample minimum detectable effect size divided by the proportion of randomized program schools that participate.

For example, findings for District B in Table 2 indicate that the minimum detectable effect size is 0.23 for average student achievement given 60 randomized schools, a 50/50 program/control group mix, and 60 third graders per school. If three of the 30 program schools did not participate, the participation rate would be 0.90. Dividing 0.23 by 0.90 yields a minimum detectable effect size of 0.26 for the adjusted impact estimator.¹⁹

¹⁹To apply this adjustment one should adopt a conservative definition of non-participation by using a threshold of little or no contact with the program. This will increase the likelihood that nonparticipating schools experienced no or negligible Reading First impacts.

Case #2: If a school that is randomized to Reading First or the control group drops-out of the evaluation, it cannot be included in the impact analysis. This can create two major problems: (1) reduced precision due to fewer schools, and (2) selection bias due to non-random omission of schools. Although precision is always an important issue, the bias created by this form of attrition is potentially a much bigger problem. The best and only truly effective defense against this threat is a strong offense created by carefully choosing sites that will want to and be able to stick with the evaluation, and closely working with sites throughout the evaluation to facilitate and ensure their continued participation.

Case #3: National statistics indicate that only half of all students who enter kindergarten at a school are there in third grade.²⁰ This high level of student mobility represents a programmatic challenge to the efficacy of Reading First because it reduces student exposure (unless they transfer among schools with the program). In addition this mobility represents a methodological challenge to the reliability and validity of the Reading First evaluation because it produces student attrition.

If Reading First does not induce, impede or otherwise appreciably modify student mobility, then it is relatively simple to deal with the methodological implications of this phenomenon. However this requires distinguishing between alternative populations for which to measure program impacts and alternative perspectives from which to do so.

From a *cross-sectional perspective* one could measure Reading First impacts on the achievement of all third graders at program schools by comparing their test scores with those of all third graders at control schools and accounting for differences in average school baseline scores. This impact would represent all current students regardless of when they came and where they came from. Thus, it would reflect an amalgam of exposures to Reading First created by the normal ongoing process of student mobility. Such impact estimates would be unbiased and relatively straightforward to interpret. In addition, they would have full-sample precision. Furthermore, they would help to contain data collection costs because they would not require tracking students who leave.

From a *longitudinal perspective* one could identify an entering kindergarten cohort, follow them over time as they progress to third grade, and measure program impacts on some facet or facets of their achievement during this period. In the presence of mobility it is useful to distinguish among impacts for the half of all students who will stay in the same school (stayers), the other half who will not (movers), and both halves together (the full cohort).

²⁰Unpublished estimates from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) kindergarten, first and third grade data.

If Reading First does not affect mobility during the evaluation, then one can readily obtain internally valid impact estimates for each of these different populations (even if mobility is high), although their data collection costs will vary. Furthermore because movers and stayers will be well represented at all sample schools, their precision will be similar to that for the full cohort.

Case #4: If Reading First changes student mobility by very much during the evaluation then things will get much more complicated. From a longitudinal perspective this is because stayers (movers) from program schools may no longer be comparable to stayers (movers) from control schools. Hence, their comparisons may produce selection bias, although they will not lose precision. The only way to avoid this bias experimentally is to follow-up an entire kindergarten cohort or a probability sample of its members regardless of where they attend school subsequently. If this were done, which might be costly, validity and precision could be maintained.

From a cross-sectional perspective it would be valid to compare all current third graders from program schools with those from control schools in order to measure program impacts on *school performance*. In addition, these comparisons would have full-sample precision. However, it would be difficult to interpret them because they would represent an unknown combination of two very different effects: (1) program impacts on the types of students who stay at or come to a school (“mobility effects”), and (2) program impacts on the achievement of students who would be at a school with or without the program (“achievement effects”). Unfortunately, there is no fully experimental (and thus internally valid) way to separate these two effects.

Promoting Generalizability

Given the real world imperatives under which the Reading First evaluation sample must be recruited, there is no way to separate considerations of precision from considerations of generalizability. Therefore, what is required is a recruitment and selection process that balances strategic preferences against opportunistic considerations and allows the relative weight of these factors to be assessed iteratively over time.

The *strategic* element of this process (its guidance system for achieving the substantive goals of the impact study) should reflect priorities that the Department of Education sets for questions that it most wants to address given limited time and resources. Thus, we will continue to work closely with the Department to ascertain its priorities on issues such as the characteristics of students, schools and settings that it most wants to represent in the study and the characteristics of schools and students that it most want to use as a basis for subgroup analyses. We will apply our knowledge of statistical requirements for addressing these questions to help launch the recruitment process in the right direction and engage the Department and others in discussions about these strategic priorities to keep this process on target as it proceeds.

The *opportunistic* element of the site recruitment and selection process (its information system for learning about and assessing fruitful targets of engagement) will require accurate and timely knowledge about states and school districts that are willing and able to participate in the evaluation, why they are willing to do so, the conditions under which they are willing to do so, and why others are not willing to do so.

The *iterative* element of the process (its feedback system for assessing and reassessing what has been accomplished and what has been learned about what else is necessary and possible) is a reflection of the fact that sample selection must unfold over time and will need to adapt continually to conditions in the field. Hence, there is no rote way to adopt standard linear textbook sampling procedures that specify a list of the total population to be sampled, adopt an algorithm for choosing elements from this sampling frame, use a random number generator to execute the algorithm, and notify sites selected.

In theory, the ideal sample for a Reading First evaluation is one that is nationally representative of the students, schools and settings eligible to receive Reading First funding. In addition, this sample should adequately represent all policy-relevant subgroups within the program population. In practice, however, the best feasible sample will be one based on tradeoffs between: political considerations (reflecting which states, districts and schools will and can participate), logistical considerations (reflecting the number of different organizations and locations that it is possible to work with and in), and statistical considerations (reflecting how the generalizability of the sample is affected by tradeoffs that are made). Hence, a more realistic goal for the study sample is to include groups of schools and students that generally are typical of those receiving Reading First funds. Findings based on this sample will thus be policy relevant even if they are not nationally representative in a narrow statistical sense.

Implications of this goal for the structure of the evaluation sample can be illustrated as follows. Assume, based on the preceding discussion, that the sample will have 60 schools in total, half of which are randomized to Reading First and half of which are randomized to a control group. At one extreme is a sample design comprising a single randomized pair of Reading First and control schools from each of 30 school districts located across the country. From a statistical perspective this design is most likely to be most representative of the broadest range of situations. From an operational perspective, however, the design would be the most difficult to implement because it would require successful negotiations with the largest number of organizational entities. It would also be the most costly and complex design to launch and maintain in the field.

At the opposite extreme would be a sample of 60 schools concentrated in one or two very large districts. From a statistical perspective, this design is the most narrowly concentrated and the least likely to reflect the range of situations that exist in the national Reading First program. However, from an operational perspective this design probably would be the easiest and least costly to implement.

The actual sample design for Reading First should be somewhere between these two extremes and should reflect the following factors.

The Allocation of Reading First Funds: Other things equal, it is important to focus the evaluation sample on states, districts and schools that represent a substantial allocation of Reading First funds. This would focus the analysis on the question: “What is the typical impact of Reading First funds that have been expended?” To date, the allocation of program dollars to states has been determined, although each state must submit a successful application to receive their allocation. However, it is not yet known how all states will allocate funds to districts. Nevertheless, a proxy that we could use for future district allocations might be constructed from data on recent allocations of Title I funds for No Child Left Behind.

Regions of the U.S. For political reasons and for reasons of face validity it is important for the evaluation sample to represent different regions of the country. Thus, it might be useful to consider approaches that stratify the U.S. into several areas (perhaps the Northeast, Southeast, Midwest, Southwest and Northwest) and aggressively recruit schools from each.

Urban, Suburban and Rural Environments: Because these environments reflect pronounced differences in economic circumstances, student and family characteristics, teacher characteristics and qualifications, administrative structures and arrangements, school resources, and local cultures, it is important to reflect these differences in the Reading First sample. This will facilitate assessing how the program functions in a wide range of settings (its robustness).

Student Race and Ethnicity. No Child Left Behind and Reading First have as a central goal narrowing the gap between the achievement of students of color and that of others. As a result, it will be critical for the evaluation sample to reflect variation in the distribution of racial and ethnic groups within and across schools.

English Language Learners. Learning to read in English is different for students who are fluent in the language of instruction than those who are not. Given the high rate of immigration to the U.S., especially from Spanish speaking countries, the number of elementary school students who do not speak fluent English has been rising over time and is likely to continue to do so. Thus, it is important, if possible to determine how Reading First impacts do or do not differ for English language learners and fluent English speakers.

Clearly there are many factors that should be considered when trying to ensure a diverse and therefore broadly generalizable sample of schools. Equally clearly, the practical realities of recruiting this sample under constraints imposed by the Reading First funding process and the institutional structure of public education in this country will create difficult tradeoffs among these considerations.

Promoting a Strong Treatment Contrast

In addition to promoting precision and generalizability, site recruitment and selection for the Reading First evaluation must help to ensure a fair test of the program by promoting a strong treatment contrast through a marked difference between reading instruction in program schools and reading instruction in control schools. There are two sides to this coin: maximizing implementation of Reading First in program schools and minimizing its emulation in control schools.

Meeting the first condition requires that schools randomly assigned to the program group implement Reading First with fidelity to its intended goals, components, and operating strategies. Unless this is done, the evaluation will produce a poor reflection of what the program can achieve because it will not represent much of a contrast with existing conditions and practices.

Hence, it will be important to recruit districts and schools that are willing and able to use Reading First funds to implement and support evidence-based early reading programs. This will help to ensure that program schools implement and sustain reforms that meet specifications laid out in the Reading First legislation and administrative guidelines.

Furthermore, it will be important to choose districts that are willing to concentrate Reading First funds in a limited number of schools to create a critical mass of resources for educational change. Such districts must be willing to withstand political forces that may force others to give something to everyone and thereby spread program resources too thinly to make a difference.

Meeting the second condition requires that participating districts not provide extra funding for the same set of services and supports in control schools during the evaluation. For example, one could imagine a scenario in which allocation of Reading First funds to program schools enables a district to redirect other resources for the same purposes to control schools, thereby reducing the observable “value-added” of Reading First. It is important therefore to recruit districts that understand the issues at stake in this regard and are willing to cooperate fully.

Preliminary Recommendations and Next Steps

Based on the issues discussed in this paper plus our past experience selecting sites for random assignment experiments, our knowledge of the Reading First legislation, and our discussions with Department staff, we make the following preliminary recommendations for a Reading First evaluation sample.

1. Recruit at least 60 schools (and preferably more to be conservative) from five states with two districts per state.
2. Include at least six schools per district and randomly assign half to Reading First and half to a control group.
3. Only choose districts with adequate baseline achievement data.
4. Focus on districts where an impact study is most likely to reflect a fair test of Reading First by providing a strong treatment contrast.
5. When preparing a sample recruitment plan, establish priorities for student, school and district characteristics to be represented.

Our next steps will be to continue meeting regularly with Department staff to develop a comprehensive and coordinated sample recruitment strategy and produce a database on the status of state Reading First applications and district allocation procedures. We also will construct a database that includes relevant information on potential districts through the “Common Core of Data” and state or district web sites and public documents. This will allow us to customize interview protocols for candidate states and districts that we contact.

The profile to be constructed for potential target districts would include data on: their numbers of eligible schools, their student characteristics, their school characteristics, their Reading Excellence programs including any other existing programs for low achieving schools, their previous student performance on state or district standardized tests, their numbers of teachers in grades K-3, the expected sizes of their Reading First grants per school and their participation in earlier reading initiatives. The information we gather from this effort will provide the basis for a concrete site recruitment plan.

Appendix A

Computing Minimum Detectable Effect Sizes

This appendix describes how minimum detectable effect sizes were computed for the alternative sample designs assessed in Tables 1–5. These findings build on theoretical work by Raudenbush (1997) and utilize empirical results from one urban school district studied by Bloom et al. (1999) plus additional empirical results produced for the present paper from data for two other urban School districts.

Introduction

The present paper examines the likely precision of alternative sample designs for estimating Reading First impacts on three student outcomes: (1) average third grade reading achievement, (2) the percentage of third graders who reach an achievement threshold, and (3) the average slope of first-to-third-grade reading achievement trajectories. The paper also examines likely precision for one teacher outcome—average survey reports about local instruction. The basic approach for estimating precision was the same in all cases. We first separated the variance in student or teacher outcomes into two components: (1) the variance across students or teachers within schools (the student or teacher variance component) and (2) the variance of mean outcomes across schools (the school variance component). We repeated this analysis with and without a baseline covariate to estimate conditional and unconditional variance components, respectively. We then used the variance component estimates to predict likely minimum detectable effect sizes for different sample designs.

To simplify our analysis, without losing much generality, it is formulated in terms of J randomized schools with n third graders per school. Proportion P of the schools are randomized to Reading First and proportion $(1-P)$ are randomized to a control group. The unconditional variance of the student or teacher outcome within schools is designated σ^2 , the unconditional variance of mean outcomes between schools is designated τ^2 , and the underlying variance of true program impacts is designated ω^2 . The conditional student or teacher variance is designated σ_*^2 and the conditional school variance is designated τ_*^2 .

J , n and P are experimental design choices, σ^2 , σ_*^2 , τ^2 and τ_*^2 are features of the schools, students and teachers being studied and ω^2 reflects how Reading First interacts with schools, teachers and students. Hence, the analyses in this paper explore the statistical precision resulting from alternative choices for J , n and P given the best available estimates of the other parameters.

We represent the statistical precision of an impact estimator, \hat{I} , as its minimum detectable effect size, $MDES(\hat{I})$. This measure is a multiple $(t_\alpha + t_{1-\beta})$ of the standard error of the impact estimator, $SE(\hat{I})$, divided by the population standard deviation of individual outcomes $\sqrt{\tau^2 + \sigma^2}$. In symbols,

$$MDES(\hat{I}) = \frac{(t_{\alpha} + t_{1-\beta})SE(\hat{I})}{\sqrt{\tau^2 + \sigma^2}} \quad (\text{A1})$$

To compute a minimum detectable effect size thus requires estimating the relevant standard error, determining the relevant multiplier, and applying the relevant population standard deviation.

Estimating the Standard Error

The impact estimator, \hat{I} , equals the difference in mean outcomes for the program and control groups, or

$$\hat{I} = \bar{Y}_p - \bar{Y}_c \quad (\text{A2})$$

Because these groups are independent,

$$VAR(\hat{I}) = VAR(\bar{Y}_p) + VAR(\bar{Y}_c) \quad (\text{A3})$$

Extending the findings of Raudenbush (1997) to incorporate a program group proportion P and true impact variation, ω^2 yields,²¹

$$VAR(\bar{Y}_p) = \frac{\omega^2}{PJ} + \frac{\tau^2}{PJ} + \frac{\sigma^2}{PJn} \quad (\text{A4})$$

and

$$VAR(\bar{Y}_c) = \frac{\tau^2}{(1-P)J} + \frac{\sigma^2}{(1-P)Jn} \quad (\text{A5})$$

Substituting Equations A4 and A5 into Equation A3 and simplifying terms yields,

$$VAR(\hat{I}) = \frac{\omega^2}{PJ} + \frac{\tau^2}{P(1-P)J} + \frac{\sigma^2}{P(1-P)Jn} \quad (\text{A6})$$

²¹ For simplicity, the present discussion omits Raudenbush's (1997) adjustment factor $\sqrt{1 + \frac{1}{J-4}}$ for a school-level covariate and $\sqrt{1 + \frac{1}{Jn-4}}$ for a student-level covariate because their values are very close to one, and hence their effect is negligible for the numbers of schools that we consider ($J \geq 10$).

Therefore,

$$SE(\hat{I}) = \sqrt{\frac{\omega^2}{PJ} + \frac{\tau_*^2}{P(1-P)J} + \frac{\sigma_*^2}{P(1-P)Jn}} \quad (\text{A7})$$

To compute this standard error for a given J, n and P requires information about ω^2 , τ^2 , τ_*^2 , σ^2 and σ_*^2 .

Table A1 reports values for σ^2 , σ_*^2 , τ^2 and τ_*^2 estimated by Bloom et al. (1999) from standardized reading test scores for 2299 third graders from two annual cohorts at 25 schools in Rochester, New York (school District A).

Table A.1
Student and School Variance Components
for Third Grade Reading Scores
(for District A)

Covariate	1991 Results	1992 Results	Mean Results
No covariate			
σ^2	101.9	105.4	103.7
τ^2	21.9	17.4	19.7
Each student's 2nd grade score			
σ_*^2	52.7	47.5	50.1
τ_*^2	6.3	4.4	5.4
Each student's 1st grade score			
σ_*^2	70.7	71.4	71.1
τ_*^2	4.2	6.0	5.1
Each school's mean 3rd grade score one year prior			
σ_*^2	104.4	106.6	105.5
τ_*^2	5.9	4.2	5.1
Each school's mean 3rd grade score two years prior			
σ_*^2	103.3	105.7	104.5
τ_*^2	4.9	2.8	3.9

SOURCE: Computations from Bloom et al. (1999).

Note that both school-level and student-level baseline performance covariates lagged either one year/grade or two years/grades reduce the school variance component dramatically. Without a baseline covariate the mean school variance is 19.7 whereas with a baseline covariate it ranges from 3.9 to 5.4. In addition, the student-level baseline covariate (but not the school-level covariate) reduces the student variance component from 103.7 without a covariate to 50.1 or 71.1 with one.

The final parameter, ω^2 was guesstimated from the thought experiment illustrated by Figure A1. This exercise reflects a liberal assumption about the likely variation in true Reading First impacts (it assumes substantial variation). Thus, we expect sample size findings that flow from the assumption to be somewhat conservative.

The figure illustrates a situation in which the true impacts of Reading First (measured as effect sizes) are normally distributed across schools in the program. The mean of this distribution is 0.5 (a sizable impact for a school-wide program) and 95 percent of all true effect sizes are between 0.0 (no impact) and 1.0 (an extremely large impact). Hence, the standard deviation of true impacts, $sd(I)$, is one fourth of the distance between zero and one, or 0.25. Thus the standard deviation of true impacts equals $0.25\sqrt{\tau^2 + \sigma^2}$ and the variance of true impacts, ω^2 equals $0.0625[\tau^2 + \sigma^2]$.

We used the preceding values for σ^2 , σ^{*2} , τ^2 , τ^{*2} and ω^2 to estimate standard errors for different values of J, n and P given Equation A7. In addition, as described below, we used corresponding values for these parameters based on several student outcome measures for a second urban school district (District B) and several teacher outcome measures for a third urban district (District C).

Determining The Multiplier

Figure A2 illustrates why the minimum detectable effect of a program impact estimator is a multiple of its standard error equal to the t value for statistical significance (t_α for a one-tail test or $t_{\alpha/2}$ for a two-tail test) plus the t value for statistical power, t_{1-B} . Hence, the multiplier equals $t_\alpha + t_{1-B}$ for a one-tail test or $t_{\alpha/2} + t_{1-B}$ for a two-tail test.

An Example

Consider an example comprising 60 schools (J) with 60 students per grade each (n) randomized half to Reading First (P=0.5) and half to a control group. Assume that each school's mean third grade reading scores lagged by two years is used as a baseline covariate and the variance components in Table A1 hold. Thus τ^2 equals 19.7, σ^2 equals 103.7, τ^{*2} equals 3.9 and σ^{*2} equals 104.5. Also assume that ω^2 equals $0.0625[\tau^2 + \sigma^2]$ as argued above and therefore is 7.7.

Substituting these parameter values into Equation A7 yields a standard error of 0.785. With 57 degrees of freedom for a one-tail t test of statistical significance at the

0.05 level given 80 percent statistical power, the value of t_{α} is 1.67 and the value of t_{1-B} is 0.85. Hence, the minimum detectable effect multiplier is 2.52.

Substituting a value of 0.785 for the standard error of the impact estimator, a value of 2.52 for the minimum detectable effect multiplier and 19.7 and 103.7 for τ^2 and σ^2 into Equation A1 yields a minimum detectable effect size of 0.18. Hence according to data from District A this sample design should be able to detect a true average program impact equal to or greater than 0.18 standard deviations.

Estimating Other Variance Components

In addition to our analysis for third grade mean achievement using data for District A we replicated this analysis using data for District B. We also used data for District B to extend our analysis to two other achievement measures: (1) the proportion (or percentage) of local third graders scoring at or above the national 50th percentile on the Stanford Achievement Test (SAT9), and (2) the slopes of student trajectories on this test for grades one, two and three, measured in normal curve equivalents per year.²² In addition, we conducted a precision analysis for measures of teacher perceptions about local instruction based on survey data from District C. Tables A2 and A3 present the variance component estimates from these analyses.

Table A.2
Estimated Variance Components
for Student Achievement Measures
(from District B)

Achievement Measure	Unconditional Variance Component for		Conditional Variance Component for	
	Students	Schools	Students	Schools
Mean third grade reading score in NCEs	276.3	69.1	276.3	27.0
Proportion of local third graders at or above the 50 th national percentile	0.220	0.034	0.220	0.017
Mean first-to-third grade slope in NCEs per year	93.6	10.5	none	None

SOURCE: Findings are based on SAT9 reading scores for 11,097 third graders from two annual cohorts at 131 schools. Conditional values control for a school-level covariate lagged two years.

²²We recognize that achievement trajectories on norm-referenced tests have limited interpretations. However, the present data were the only relevant information that was readily available and they might well yield a reasonable approximation to the precision one can expect for estimating program impacts on student growth curves.

Table A.3

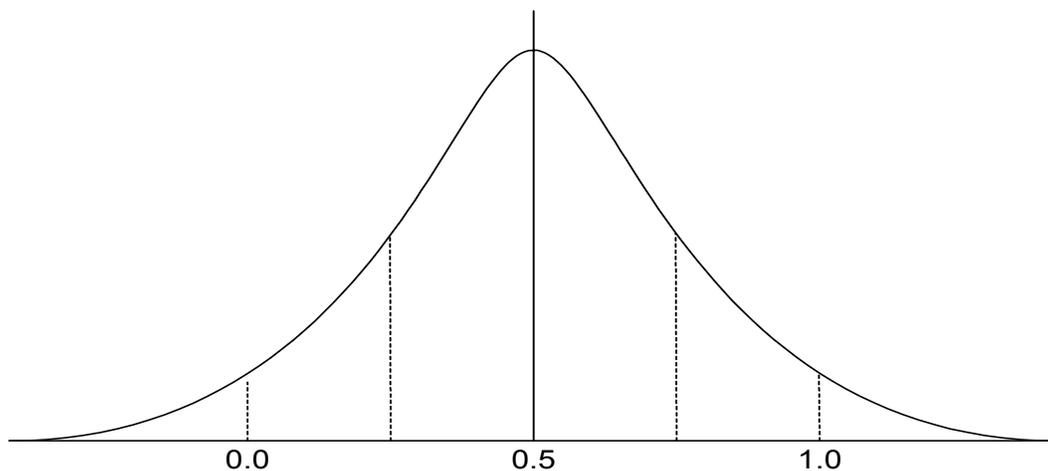
**Estimated Variance Components
for Teacher Reports on Instruction
(from District C)**

Instruction Measure	Unconditional Variance Component for		Conditional Variance Component for	
	Teachers	Schools	Teachers	Schools
Curriculum appropriateness	0.220	0.049	0.220	0.017
Teacher development	0.282	0.084	0.282	0.014
Teacher efficacy	0.212	0.033	0.212	0.025
MEAN	0.238	0.055	0.238	0.019

SOURCE: Findings are based on surveys of 220 teachers from seven elementary schools for two years. Conditional values control for school-level means for each measure lagged one year.

Figure A1

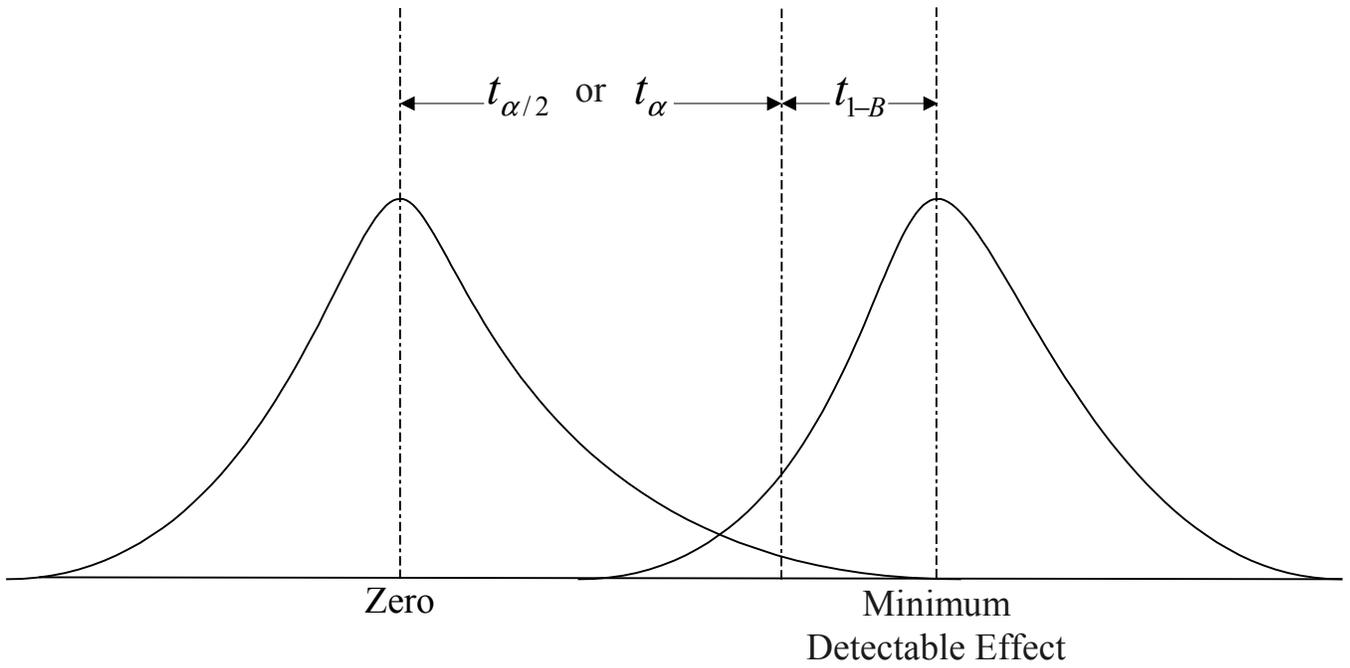
**Assumed Distribution Across Schools of the
True Effect Sizes of Reading First**



Effect Size (in standard deviations)

Figure A2

The Minimum Detectable Effect Multiplier



$$\text{One-Tail Multiplier} = t_{\alpha} + t_{1-B}$$

$$\text{Two-Tail Multiplier} = t_{\alpha/2} + t_{1-B}$$

References

- Bloom, Howard S. Forthcoming. "Using Cluster Randomization with Nonexperimental Methods for Place-Based Evaluation." New York: MDRC.
- Bloom, Howard S. and James A. Riccio. 2002. "Using Place-Based Random Assignment and Comparative Interrupted Time-Series Analysis to Evaluate a Community Employment Program for Public Housing Residents: The Jobs-Plus Demonstration." (Prepared for the Campbell Collaborative Conference on Place-Based Randomized Trials in Bellagio, Italy.) New York: MDRC.
- Bloom, Howard S., Carolyn J. Hill and James Riccio. 2001. "Modeling the Performance of Welfare-to-Work Programs: The Effects of Program Management and Services, Economic Environment, and Client Characteristics." New York: MDRC.
- 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: 445-469.
- Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review*. Vol. 19 No. 5: 547-556.
- Bloom, Howard S. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review*. Vol. 8 No. 2, 225-246.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Science*, 2nd edition. Hillsdale, NJ: Lawrence Erlbaum.
- Finn, Jeremy D. and Charles M. Achilles. 1999. "Tennessee's Class Size Study: Findings, Implications, Misconceptions." *Educational Evaluation and Policy Analysis*. Vol. 21 No. 2: 97-109.
- Gail, Mitchell H., Steven D. Mark, Raymond J. Carroll, Sylvan B. Green, and David Pee. 1996. "On Design Considerations and Randomization-Based Inference for Community Intervention Trials." *Statistics in Medicine*. Vol. 15: 1069-1092.
- Lipsey, Mark. 1990. *Design Sensitivity: Statistical Power for Experimental Research*, pp. 51-56. Newbury Park, CA: Sage Publications.
- Nye, Barbara, Larry V. Hedges, and Spyros Konstantopoulos. 1999. "The Long-Term Effects of Small Classes: A Five-Year Follow-up of the Tennessee Class Size Experiment." *Educational Evaluation and Policy Analysis*. Vol. 21 No. 2: 127-142.
- Raudenbush, Stephen W. 1997. "Statistical Analysis and Optimal Design in Cluster Randomized Trials." *Psychological Methods*. Vol. 2 No. 2: 173-185.