Measuring the Impacts of Whole-School Reforms: Methodological Lessons from an Evaluation of Accelerated Schools

Howard S. Bloom
Manpower Demonstration Research Corporation
New York, N.Y.

This work was supported by the Planning and Evaluation Service of the U.S. Department of Education, the Pew Charitable Trusts, and the Alfred P. Sloan Foundation.

The findings and conclusions presented in the paper do not necessarily represent the official positions or policies of the funders.

An earlier version of this paper was prepared for the annual meeting of the American Educational Research Association held in Seattle, Washington, in April 2001 and was presented as the first Jerry Miner Lecture at the Maxwell School of Citizenship and Public Affairs of Syracuse University in April 2001.

The author would like to thank Fred Doolittle, James Kemple, Jeffery Rodamar, Leanna Steifel, and Johnny Yinger for their helpful comments on the issues addressed in the paper. The author also would like to acknowledge the contributions of JoAnn Rock, who had the original idea to use interrupted time-series analysis to evaluate Accelerated Schools and was project director and principal impact analyst for the Accelerated Schools study.

Abstract

The goal of the present paper is to introduce education researchers to a new approach for measuring the impacts of whole-school reforms. The approach is based on interrupted time-series analysis, which has been used to evaluate programs in many fields but has not been used widely to study education initiatives. The application presented measures program impacts on student performance by comparing standardized test scores for a number of annual student cohorts in a specific grade after a reform is launched (its follow-up period) with the scores of cohorts from several years before the reform was launched (its baseline period). The approach is used to measure impacts on three facets of student performance: (1) average (mean) test scores, which summarize impacts on total performance; (2) the distribution of scores across specific ranges, which helps to identify where in the distribution of student performance impacts were experienced; and (3) the variation (standard deviation) of scores, which indicates how the disparity in student performance was affected. To help researchers use the approach, the paper lays out its conceptual rationale, describes its statistical procedures, explains how to interpret its findings, indicates its strengths and limitations and illustrates how it was used to evaluate a major whole-school reform model—Accelerated Schools.

Contents

This work was supported by the Planning and Evaluation Service of the U.S. Department of Education, the Pew Charitable Trusts, and the Alfred P. Sloan Foundation.

The findings and conclusions presented in the paper do not necessarily represent the official positions or policies of the funders.

An earlier version of this paper was prepared for the annual meeting of the American Educational Research Association held in Seattle, Washington, in April 2001 and was presented as the first Jerry Miner Lecture at the Maxwell School of Citizenship and Public Affairs of Syracuse University in April 2001.

The author would like to thank Fred Doolittle, James Kemple, Jeffery Rodamar, Leanna Steifel, and Johnny Yinger for their helpful comments on the issues addressed in the paper. The author also would like to acknowledge the contributions of JoAnn Rock, who had the original idea to use interrupted time-series analysis to evaluate Accelerated Schools and was project director and principal impact analyst for the Accelerated Schools study.

1. Introduction

This paper presents a new approach for estimating the impacts of whole-school reforms and describes how it was used for a recent evaluation of the Accelerated Schools reform model (Bloom et al., 2001). The approach is an application of interrupted time-series analysis, which has been used to evaluate programs in many fields but has not been used widely for education evaluations.[1] Interrupted time-series analysis is one component of a more comprehensive quasi-experimental strategy that is being developed by the Manpower Demonstration Research Corporation (MDRC) to evaluate school reforms. Thus, the present paper is a first step toward this end.

The analyses described below address three main evaluation questions:

1. How, if at all, did a specific reform affect average student performance?

2. How, if at all, did it affect the variation in student performance?

3. How, if at all, did it affect the distribution of student performance?

The first question focuses on the overall *effectiveness* of a reform by measuring its total improvement per student. The second question focuses on the *equity* of a reform by measuring the degree to which it increases or decreases the disparity in student performance. The third question helps to identify how different types of students (those who are below average locally, about average locally, or above average locally) are affected by the reform.

The present approach to addressing these questions has four main features:

1. The approach *projects* what student performance for a specific grade in a school would have been without the reform during a multiyear follow-up period after the reform was launched. This projection is based on the pattern of student performance for that grade during a multiyear baseline period before the reform was launched. The difference between actual and projected student performance provides an estimate of the impact of the reform—the change in performance that it caused.

2. The approach *controls for* observed changes over time in selected student background characteristics by integrating a multiple regression model with the baseline projection model.

[1] Shadish, Cook and Campbell (forthcoming) provide a comprehensive review of the interrupted time-series literature. Campbell and Stanley (1966) and Cook and Campbell (1979) are the most widely read sources on the topic. Bloom (1999) describes how to use interrupted time-series analysis to measure program impacts on student performance.

3. The approach *accounts for* year-to-year fluctuations in student performance due to "cohort effects" by integrating a variance component model with the regression model and the baseline projection model.

4. The approach *combines* impact findings for different schools that use different tests to measure student performance.[2]

Section 2 of the paper discusses each major feature of the impact estimation approach and Sections 3 through 5 illustrate how it was applied for the evaluation of Accelerated Schools. Section 6 concludes with a discussion of next steps in the development of a more comprehensive evaluation strategy.

[2] Each school must use the same test over time, however.

2. The Approach

This section describes the main features of the current impact estimation approach and outlines the issues, options and decision criteria to be considered when using it.

2.1 Projecting the "Counterfactual"

The core of the strategy is an interrupted time-series analysis that measures the impact of a reform as the subsequent *deviation from the past pattern of student performance* for a specific grade. Hence, the analysis compares the performance of different student cohorts for a given grade, although for completeness it should be replicated for as many grades as possible.

To apply the approach for a given grade requires a consistent measure of student performance (usually a standardized test) for the grade during a number of baseline years before the reform was launched. Furthermore, because whole-school reforms usually take three to five years to implement, a fair test of their impacts requires maintaining the same baseline performance measure (test) for at least three to five follow-up years.

The key to success for any method of estimating the impacts of a school reform is its ability to project what student performance would have been in the absence of the reform. This hypothetical state of the world is usually referred to as a "counterfactual." Only by comparing actual student performance to this counterfactual can one obtain valid estimates of the reform's impacts.

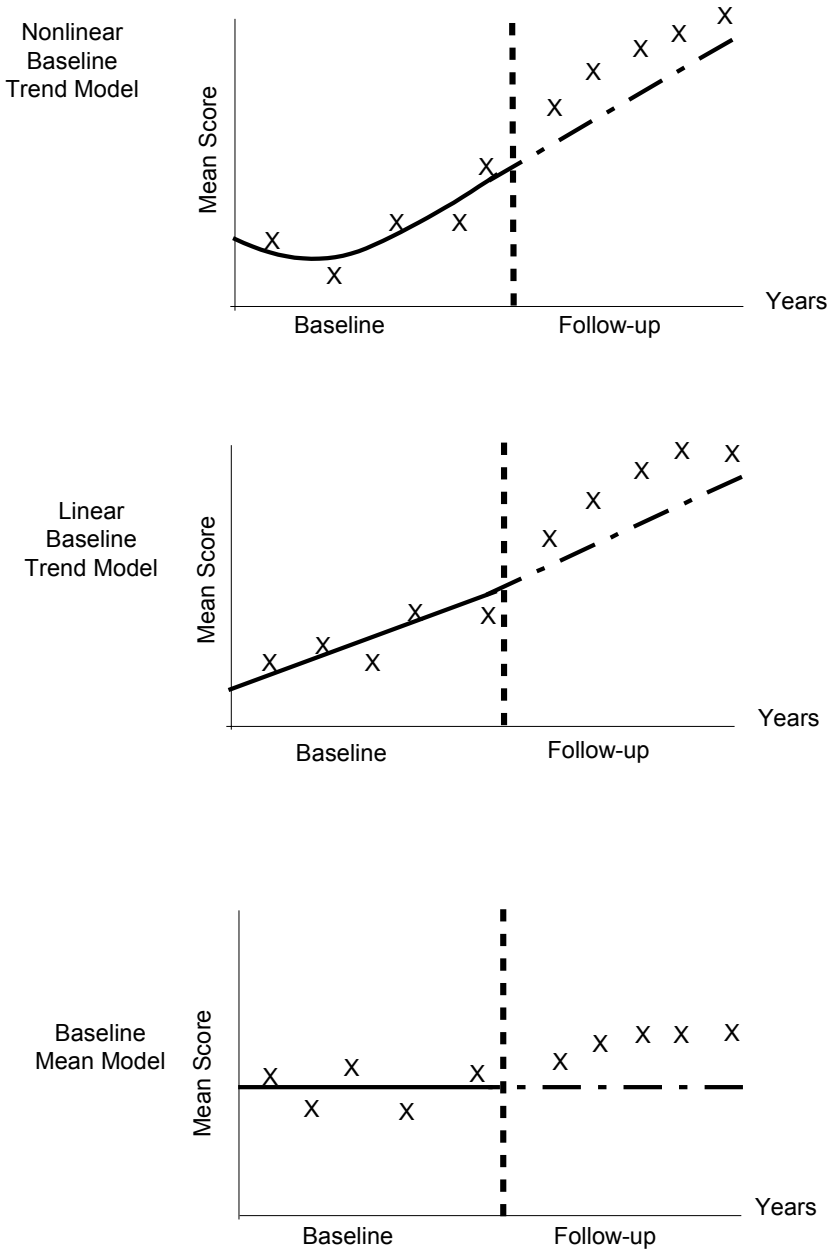
An interrupted time-series approach to projecting a counterfactual proceeds from two related premises: (1) that past experience is the best predictor of future experience in the absence of systemic change, and (2) that multiple observations of past experience predict future experience better than a single observation. However, different baseline patterns may suggest different projection models, as illustrated by Figure 1.

- ***A nonlinear baseline trend model***. The top graph in the figure illustrates baseline test scores for annual student cohorts in a specific grade that are rising at an increasing rate.[3] To project this pattern into the follow-up period requires a nonlinear model. However, to use such a model requires estimating at least three temporal parameters, which is difficult to do with only four or five years of baseline test data, and thus only one or two temporal degrees of freedom.[4] Furthermore, given how often schools change their standardized tests, very few of them maintain even five years of consistent baseline data. Thus, nonlinear projection models are probably not feasible in practice.

[3] In theory, there are many potential nonlinear patterns of scores that might rise or decline over time at an increasing or decreasing rate.

[4] Note that what limits the precision of these parameter estimates is the number of baseline years for which consistent test data are available, not the number of students tested each year.

Figure 1
Alternative Baseline Projection Models



- ***A linear baseline trend model***. The middle graph in the figure illustrates baseline test scores that are rising at a constant rate across consecutive annual student cohorts. A linear model could be used to project this pattern into the follow-up period. However, doing so requires estimating two temporal parameters (an intercept plus a slope) from data for four or five points in time and thus, only two or three temporal degrees of freedom. Hence, the precision of these estimates may be limited and the corresponding projection error may be substantial, especially for later follow-up years.

  Consider, for example, the error than can occur from overestimating a baseline slope. When this happens, future projections will overestimate the counterfactual by an increasing amount each year. While this might not be problematic for early follow-up years, before the impacts of most school reforms are expected to materialize, it could be highly misleading for later follow-up years, when such impacts are most likely to occur, if they are going to do so.

  Thus, although a linear model is intuitively appealing, it is a potentially risky way to make counterfactuals for school reforms. Therefore, one should only use this model if the observed baseline trend is highly consistent and there is good reason to believe that it will continue for four to five baseline years plus four to five follow-up years.

- ***A baseline mean model***. The bottom graph in the figure illustrates baseline test scores that vary randomly across annual student cohorts with no clear sign of a systematic increase or decrease over time. To project this pattern into the follow-up period, one could use the baseline mean score.

  This model is the simplest and least risky of the three considered. Because it uses several years of baseline experience—not just one—to project counterfactuals, it reduces the sensitivity of impact estimates to aberrations in student performance. In addition, because it does not try to use limited information about changes in performance over time to compute a constant slope (as in linear models) or a time-varying slope (as in nonlinear models), it avoids the especially large errors that can occur in later follow-up years if one attempts to estimate based on a slope and "guesses wrong."

  Furthermore, it is possible to use a baseline mean model with as few as three years of baseline test data.[5] This reduces the data requirements of the interrupted time-series approach and thereby broadens its potential range of applications.

[5] In principle, the approach could be used with only one or two years of baseline test data. However, this would markedly reduce its protection against errors due to unusual student performance or local idiosyncratic events.

Equations 1–3 below specify regression models that can be used to estimate the impacts of school reforms from the preceding three projection models. Each regression could be estimated from a pooled sample of scores for all student cohorts that were tested during the baseline or follow-up periods.

A nonlinear baseline trend model

$$Y_i = f(t_i) + \Sigma D_k FY_{ki} + e_i \qquad \textit{(1)}$$

A linear baseline trend model

$$Y_i = a + bt_i + \Sigma D_k FY_{ki} + e_i \qquad \textit{(2)}$$

A baseline mean model

$$Y_i = A + \Sigma D_k FY_{ki} + e_i \qquad \textit{(3)}$$

where:

- Y_i = the test score for student i,
- t_i = a counter for time which is zero for students in the first baseline cohort and increases by one unit for each subsequent cohort,
- $f(t_i)$ = a non-linear function of t_i,
- FY_{ki} = equals one if student i was a member of the cohort for follow-up year k and zero otherwise, and
- e_i = a random error term for student i.

Equation 1 specifies a non-linear trend, $f(t_i)$, that is estimated from baseline scores and is used to project the counterfactual for each follow-up year. The coefficient, D_k, for the student cohort in follow-up year k, represents the average *deviation* of test scores for that cohort from its projected counterfactual, $f(t_k)$. Hence, D_k, represents the reform's impact for follow-up year k. Likewise, the standard error for an estimate of D_k represents the standard error of the impact estimate for that year.

Equation 2 specifies a linear baseline trend, $(a + bt_i)$, that plays the same role as $f(t_i)$ in Equation 1. Thus, Equation 2 represents the impact of a reform as its deviation, D_k, from $(a + bt_k)$. Equation 3 specifies a baseline mean, A, that plays the same role as $(a + bt_i)$ and $f(t_i)$ in the preceding models. Thus, Equation 3 represents the impact of a reform as its deviation, D_k, from A, where A is the same for cohorts in all follow-up years.

Note that specifying a separate indicator variable, FY_k, for each follow-up year "removes" test scores for these years from estimation of the baseline model. Thus, one can jointly estimate the baseline model and subsequent deviations from its projections by pooling data for all baseline and follow-up years.

2.2 Controlling for Changes in Student Characteristics

One simple way to extend the preceding analysis is to control explicitly for systematic differences over time in the background characteristics of student cohorts. To do so, one could add individual student characteristics, X_{ji}, to the baseline projection model. Equations 4–6 present the corresponding regression models for this extension: [6]

A nonlinear baseline trend model with student characteristics

$$Y_i = f(t_i) + \Sigma D_k F Y_{ki} + \Sigma C_j X_{ji} + e_i \qquad \textit{(4)}$$

A linear baseline trend model with student characteristics

$$Y_i = a + bt_i + \Sigma D_k F Y_{ki} + \Sigma C_j X_{ji} + e_i \qquad \textit{(5)}$$

A baseline mean model with student characteristics

$$Y_i = A + \Sigma D_k F Y_{ki} + \Sigma C_j X_{ji} + e_i \qquad \textit{(6)}$$

Schools typically keep records on student characteristics such as gender, age (and thus age-for-grade), race or ethnicity, special education status, English language proficiency and eligibility for subsidized school meals. Hence, it is frequently possible to control for these characteristics in an impact analysis. For example, if a school experienced a racial transition then data on the race or ethnicity of its students could be used to control for changes in the racial composition of its subsequent student cohorts. Nevertheless, interrupted time-series analysis may not be appropriate for schools that experience radical changes in their student population because of the limited number of student background characteristics that usually are available to control for such changes, and because of the limited statistical control (for selection bias) usually provided by regression methods even when extensive background data are available.

When possible, it is especially useful to include as an X_{ji} a measure of individual student performance on a related test in an earlier grade. Such pretests (discussed in Section 6.2 as the basis for "value-added" models) are the most effective way to control for systematic differences in the composition (and thus abilities) of annual student cohorts. This is because student performance on a pretest is the best predictor of performance on a posttest—indeed far better than all other typically-available student characteristics combined.[7]

---

[6] Note that the error term, e_i, in Equations 4–6 differs from that in Equations 1–3.

[7] This can be demonstrated by comparing the explanatory power (R^2) of a regression of posttests on pretests for a group of students with the corresponding explanatory power of a regression of posttests on all other background characteristics combined. In all cases explored by the author, the explanatory power of the pretest was much greater than that of all other characteristics combined.

2.3 Accounting for Cohort Effects

Another important factor to consider when using interrupted time-series analysis to evaluate school reforms is the extent to which student performance varies from year to year due to "cohort effects." Cohort effects reflect all factors that influence the performance of students *as a group* rather than individually. For example:

- The professional attitudes and behaviors of specific teachers might vary from year to year in ways that affect their whole class. Thus, student cohorts for different years in a grade might have different experiences with the same teacher.

- Some teachers might be assigned to different grades in different years so that different student cohorts experience different teachers for the same grade.

- A small number of students in each class might "set the tone" for the year in ways which affect their entire class.

- Testing conditions (the temperature, noise level or clarity of instructions) might vary in ways that affect the measurement of student performance for each class.

For these and other reasons, test scores for a specific grade in a school might vary from year to year by more than can be explained by individual random sampling error. If so, then one must account for this variation when estimating the projection error of a counterfactual and the standard error of its corresponding impact estimate. If cohort effects exist but are ignored, the standard error of an impact estimate will be understated and its statistical significance will be overstated.

Cohort effects can be accounted-for by adding a random error term, v_t, for each cohort to the random error term, e_i, for each student. Versions of this variance component structure—called "random effects" models or "mixed models"—are used frequently to analyze panel data in econometrics (Greene, 1997, 623-632). In addition, this error structure represents the simplest form of a hierarchical linear model (Bryk and Raudenbush, 1992, 17-18). These models can be estimated using Maximum Likelihood procedures that are available from many standard software packages (for example, SAS PROC MIXED). Adding this error component structure to the impact estimation models in Equations 4–5 yields:[8]

A nonlinear baseline trend model with student characteristics and cohort effects

$$Y_i = f(t_i) + \Sigma D_k F Y_{ki} + \Sigma C_j X_{ji} + v_t + e_i \qquad \textit{(7)}$$

[8] Note that the error term, e_i, in Equations 7–9 differs from that in Equations 1–3 and 4–6.

A linear baseline trend model with student characteristics and cohort effects

$$Y_i = a + bt_i + \Sigma D_k F Y_{ki} + \Sigma C_j X_{ji} + v_t + e_i \qquad \textit{(8)}$$

A baseline mean model with student characteristics and cohort effects

$$Y_i = A + \Sigma D_k F Y_{ki} + \Sigma C_j X_{ji} + v_t + e_i \qquad \textit{(9)}$$

Thus, the total error term for each sample member has two components that jointly determine the unexplained year-to-year variation in mean test scores. To see how these components affect the standard error of an impact estimate first note that the total individual variance of the error term is:[9]

$$VAR(v_t+e_i) = VAR(v_t) + VAR(e_i)$$

$$= \tau^2 + \sigma^2 \qquad \textit{(10)}$$

where:

τ^2 = the variance of v_t across cohorts (the "cohort component"), and

σ^2 = the variance of e_i within cohorts (the "individual component").

Now recall from Figure 1 how average annual test scores varied around the baseline pattern. This "unexplained" year-to-year variation in average test scores is the source of estimated uncertainty about future projections from the baseline pattern.[10] Thus, it is also the source of estimated uncertainty about impact estimates based on these projections. The more tightly average test scores cluster around the baseline pattern, the more confidence one can place in future projections from this pattern and thus, the more confidence one can place in impact estimates from these projections.

Bloom (1999; Appendix B) illustrates that the standard error of an impact estimate obtained from an interrupted time-series analysis is directly proportional to the square root of the year-to-year unexplained variance in mean annual test scores. He also

[9] The variance of the sum of the components equals the sum of the variances of the components because the components are independently distributed random variables.

[10] *Estimated* uncertainty refers to the uncertainty that is measured by the standard error of an impact estimate. It reflects the variation in baseline test scores that is not "explained" by the baseline projection model. However, there is always additional unmeasured uncertainty about baseline projections and thus about impact estimates due to the possibility that the wrong baseline projection model was used. This uncertainty is not readily quantifiable.

demonstrates that with *n* students per cohort the unexplained variance in mean annual test scores is[11]

$$VAR(v_t + \bar{e}_t) = \tau^2 + \sigma^2/n \qquad \textit{(11)}$$

Now consider what would happen if a cohort effect existed ($\tau^2>0$) but were ignored and the impacts of a reform were estimated by an ordinary least squares (OLS) analysis from Equation 4, 5 or 6. OLS computes the unexplained variance of mean annual test scores as the total unexplained variance per student divided by the number of students per cohort[12] or $\tau^2/n + \sigma^2/n$. However, as noted above, the correct variance is $\tau^2 + \sigma^2/n$. Thus, the OLS variance is too small and the magnitude of this underestimate depends on the size of the cohort effect relative to that of the individual effect.

One way to represent the relative size of these two random effects is an intraclass correlation, ρ, where[13]

$$\rho = \tau^2/(\tau^2 + \sigma^2). \qquad (12)$$

Intuitively, ρ is the proportion of total unexplained variation in individual test scores due to cohort differences.[14]

Given this definition for ρ, the ratio of the correct variance of unexplained average test scores to its OLS counterpart equals $1 + (n-1)\rho$. Because the standard error of an impact estimate is proportional to the *square root* of the unexplained annual test score variance, the ratio of the correct standard error to its OLS counterpart is $\sqrt{1+(n-1)\rho}$. I refer to this expression as the "Cohort Effect Multiplier" (CEM).

Table 1 lists the values of this multiplier for different cohort sizes and cohort effects. Note that if there is no cohort effect ($\rho = 0$) the Cohort Effect Multiplier is one and the OLS standard error equals the correct standard error, regardless of cohort size.

---

[11] One important implication of this finding is that cohort size, n, can reduce year-to-year unexplained variation in test scores only through the individual variance component, σ^2/n. Thus, although larger schools with larger cohorts per grade may have smaller year-to-year unexplained test score variation, there are diminishing returns to school size in this regard.

[12] To simplify the discussion it assumes a constant annual cohort size, n, which is a reasonable approximation for the Accelerated Schools sample. Also to simplify the discussion it is formulated in terms of the population parameters, τ^2 and σ^2, instead of their sample counterparts, which would be used for the actual calculations. However, as the sample size increases the difference between these population parameters and their corresponding sample statistics decreases.

[13] An intraclass correlation is a common way to represent the relationship between the variance within groups and the variance between groups (Murray, 1995).

[14] Likewise, $(1-\rho)$ is the proportion of total unexplained individual variation in test scores due to random individual differences.

Table 1

The Cohort Effect Multiplier
for Different Cohort Effects and Cohort Sizes

Cohort Effect (ρ)	Cohort Size (n)				
	50	**75**	**100**	**125**	**150**
.00	1.00	1.00	1.00	1.00	1.00
.01	1.22	1.32	1.41	1.50	1.58
.02	1.41	1.57	1.73	1.87	1.99
.03	1.57	1.79	1.99	2.17	2.34
.04	1.72	1.99	2.23	2.44	2.64
.05	1.86	2.17	2.44	2.68	2.91

NOTE: The Cohort Effect Multiplier equals $\sqrt{1+(n-1)\rho}$.

However, when cohort effects exist—even small ones—the discrepancy between an OLS standard error and the correct standard error can be substantial. Consider what happens to cohorts of 100 students (four classes of 25 students each for one grade in a school). When ρ is 0.01 the Cohort Effect Multiplier is 1.41. This means that even though cohort differences account for only one percent of the unexplained variation in individual test scores, the correct standard error is almost half-again as large as its OLS counterpart. If ρ equals 0.05 the correct standard error is almost two and a half times its OLS counterpart.

Little is known about the magnitudes of cohort effects for time-series of student test scores. The only previously available estimates are those reported by Bloom (1999) based on a linear trend analysis of standardized reading and math tests for third-grade and sixth-grade students from 25 elementary schools in Rochester, N.Y. For reading, the median value of ρ for the 25 schools was less than 0.01 for third grade and sixth grade. For math, the median value was 0.02 for both grades. Findings from the Accelerated Schools evaluation reported in Section 3.4 of the present paper are consistent with these results. This suggests that cohort effects may be small but too large to ignore.

2.4 Combining Impact Findings Across Schools

The preceding sections focused on how to estimate the impacts of an education reform at a single school. This section considers how to summarize these findings for a sample of schools. The first step in this process is to express all findings in a common metric. The next step is to combine impact findings in ways that are interpretable and have meaningful statistical properties.

2.4.1 Expressing Impacts in a Common Metric

Different schools that adopt an education reform often use different tests to measure student performance. Thus, an important first step in a multi-school evaluation is to express student performance in a metric that facilitates combining findings across schools.

However, it is only possible to combine findings in a meaningful way if all schools measure the same constructs—at least to a reasonable degree of approximation. This conceptual point lies at the heart of what it means to combine impact findings. For example, it does not make sense to combine measures of impacts on completely different constructs (for example, student performance in Spanish and algebra). On the other hand, it might be quite reasonable to combine findings from two different measures of the same construct (for example, two tests of third-grade reading proficiency). Therefore the first step in combining impact findings for schools is to carefully examine how they define and measure student performance.

A second point to consider is a statistical one that lies at the heart of what information is being conveyed by a student performance measure. Cardinal measures—such as Normal Curve Equivalents (NCEs) and most scale scores for standardized reading and math tests—convey a sense of both (1) the rank *order* of individual student performance (indicating who performs better than whom) and (2) the *amount by which* performance differs between individuals (indicating by how much each student out-performs or under-performs others). Ordinal measures—such as percentiles for norm-referenced tests—convey only the first of these two types of information.

Because it is only valid to perform arithmetic operations (addition, subtraction, multiplication, division, square roots, etc.) on cardinal measures, they are required for the interrupted time-series analysis described in this paper. In addition, they are required for computing means, variances or standard deviations of impact estimates for groups of schools.[15]

If all schools in a sample measure the same basic constructs and they use cardinal measures to do so, then it is easy to express their findings in a common metric. One way to do so is to standardize the scores for each student in a given grade from a school relative to the mean and standard deviation of all *baseline* scores for that grade and school. For example, one could standardize the scores for all third-graders who were tested during the baseline and follow-up periods relative to the baseline mean and the baseline standard deviation. This would involve computing a Z-score for each student as follows:

$$Z_{ji} = \left(X_{ji} - \bar{X}j \right) / S_j \qquad \textbf{(13)}$$

[15] It is possible and often desirable, however, to express final impact findings that have been computed from cardinal performance measures in terms of their ordinal equivalents (for example, percentiles). However, this transformation must occur after all computations have been done.

where:

Z_{ji} = the Z-score for third-grade student i from school j,

X_{ji} = the scale score (NCE or other) for third-grade student i from school j,

$\bar{X}_j$ = the mean scale score for all third-grade students tested during the baseline period at school j,

S_j = the standard deviation of the scale scores for all third-grade students tested during the baseline period at school j.

Each student's Z-score reflects how far he or she was above or below the baseline mean in multiples (or fractions) of the baseline standard deviation. Thus, for example, a Z-score of 0.5 implies that a student scored half a baseline standard deviation above the baseline mean.

After Z-scores are computed for all students from a school, the impacts can be estimated as described in Sections 2.1 – 2.3 above. Because the new unit of student performance is a baseline standard deviation, the impact estimate is also expressed in this unit. Thus, an estimated impact on average student test scores of 0.25 implies a reform-induced increase in the average test score equal to 0.25 baseline standard deviations.

This impact measure is the same as an "effect size" which is used for meta-analyses to combine impact estimates across studies (Hedges and Olkin, 1985 and Cooper and Hedges, 1994). Effect size is also a common way to report impacts for evaluations of education programs.

Although judgments about whether a specific effect size is large or small are ultimately arbitrary, some guidelines do 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. Lipsey (1990) 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 are from education research. The bottom third of this distribution (small impacts) ranged from 0 to 0.32, the middle third (moderate impacts) ranged from 0.33 to 0.55, and the top third (large impacts) ranged from 0.56 to 1.26.

Another benchmark for assessing effect size estimates is provided by findings from the Tennessee elementary school class size experiment, STAR. This study is a relevant point of reference for three reasons. First, it represents a major effort to improve elementary school student performance. Second, it estimates program impacts using a randomized experiment, which is the most rigorous way to do so. Third, effect size is defined comparably for the evaluation of STAR and the types of analyses described in the present paper.

Effect size estimates for STAR ranged mainly between about 0.15 and 0.25 standard deviations (for example, see Finn and Achilles, 1999 and Nye, Hedges and

Konstantopoulos, 1999). This difference in average elementary school student performance in reading and math was produced by a class size reduction from 22-26 students per normal size class to 13-17 students per reduced size class.

As long as effect sizes are defined and constructed the same way for all schools in a sample (which will result from the Z-score transformation described above), it is appropriate to combine them across schools. However, it is often not possible to combine (or compare) effect size estimates from studies using different research designs or outcome measures (Olejnik and Algina, 2000). Thus caution should be used when combining effect sizes measures from different studies—even studies of the same education reform model.

2.4.2 Combining Impact Estimates

There are three main options for combining impact findings across schools and the choice of option depends on the population to which one is trying to generalize and the assumption that one is willing to make about how impacts vary, or not, across schools.

The Options: To help distinguish among the three options I refer to them as "constant-effect" estimators, "fixed-effect" estimators and "random-effect" estimators.

Constant-effect estimators assume that the true impacts of a reform are the same for all schools—they are homogeneous. Based on this assumption, constant-effect estimators use impact findings for a sample of schools to estimate the true constant impact for a larger population of schools.

Fixed-effect estimators assume that the true impacts of a reform may vary across schools—they are heterogeneous—and that findings for the current sample do not readily generalize to a larger identifiable population. Hence, fixed-effect estimators use impact findings for a sample of schools to estimate the mean true impact *for that sample.* In other words, they define the population of interest as the sample at hand.

Random-effect estimators assume that true impacts vary across schools and that the current sample of schools statistically represents a larger identifiable population. Thus, random-effect estimators use impact findings from a sample to estimate the mean true impact for a larger population.

The estimation procedure for each option flows directly from the statistical inference (generalization) being attempted and the assumptions being made. For example, a good constant-effect estimator would be unbiased and have a minimum error variance for estimating an impact that did not vary across schools.

These are the properties of a precision-weighted mean of individual school impact estimates, such that:

$$\hat{I}_c = \sum_{1}^{m} w_j \hat{I}_j \qquad \textbf{(14)}$$

where:

$\hat{I}_c$ = the combined estimator of the assumed constant impact,
w_j = a weight for school j that is proportional to the precision of its impact estimate and sums to one across schools in the sample,
$\hat{I}_j$ = the impact estimate for school j, and
m = the number of schools in the sample.

The weights for this estimator could be specified in terms of the sample size for each school or the estimated standard error for each school's impact estimate.

The estimation error for the combined estimate is a weighted average of the estimation error for each school. Thus, the combined standard error is:

$$\hat{SE}(I_c) = \sqrt{\sum_{1}^{m} w_j^2 \hat{VAR}(I_j)} \qquad \textbf{(15)}$$

A fixed-effect estimator of the true mean impact for a sample of schools can be obtained from the mean of their impact estimates, weighting each school equally, where:

$$\hat{I}_f = \frac{\sum_{1}^{m} \hat{I}_j}{m} \qquad \textbf{(16)}$$

The standard error of this estimator is

$$\hat{SE}(I_f) = \frac{\sqrt{\sum_{1}^{m} \hat{VAR}(I_j)}}{m} \qquad \textbf{(17)}$$

Thus, its total error is proportional to the total error in the school-specific impact estimates for the sample.

A random-effect estimator is computed in the same way as a constant-effect estimator, so that:

$$\hat{I}_r = \hat{I}_c = \sum_{1}^{m} w_j \hat{I}_j \qquad \textbf{(18)}$$

However, the standard error of a random-effect estimator differs from that for a constant-effect estimator in a way that reflects the fundamental difference between their underlying assumptions. Specifically, random-effect estimators relax the assumption that impacts are the same for all schools and allow them to vary in response to differences in local circumstances. Therefore, random-effect estimators have two sources of uncertainty that must be reflected in their standard errors: (1) error in each school's impact estimate *plus* (2) differences in their true impacts.

The first source of uncertainty is the same as that for constant-effect estimators.[16] The second source reflects the extent to which true mean impacts can vary from one potential sample of schools to another.[17] Thus, the standard error of a random-effect estimator equals the square root of the sum of the error variance of a constant effect estimator *plus* the actual variance in true impact across schools divided by the number of schools, or

$$SE(\hat{I}_r) = \sqrt{VAR(\hat{I}_c) + \frac{VAR(I_j)}{m}}$$

$$= \sqrt{\sum_{1}^{m} w_j^2 VAR(\hat{I}_j) + \frac{VAR(I_j)}{m}} \qquad \textbf{(19)}$$

Choosing Among the Options. When thinking about which of the preceding estimators to use for an evaluation it is important to consider: (1) the breadth of the generalization that is desired, (2) the realism of the assumptions required for this generalization, (3) the statistical power of the inference involved, and (4) the size and composition of the sample of schools upon which it is based.

Constant-effect estimators have the smallest standard errors and thus, the greatest statistical power of the three preceding options. In addition, they generalize impact findings beyond the sample of schools observed. However, their fundamental assumption of homogeneous impacts is unduly restrictive and

[16] This first source of error is also the only source of error for fixed-effect estimators. Thus, the only difference between the standard errors for constant-effect estimators and those for fixed-effect estimators is that owing to differences in how they weight impact estimates for each school.

[17] This error is defined in terms of the *sampling distribution* of the combined estimator, which is a theoretical construct that represents all possible estimates that might have been obtained from all possible samples of schools that might have been drawn from the population of interest.

inconsistent with the reality of how constraints and opportunities differ among schools that implement an education reform. Because of this, *constant-effect estimators probably are not appropriate for evaluating school reforms.*

Fixed-effect estimators are realistic with respect to the likelihood that true impacts vary across schools. In addition, they are realistic with respect to the non-probabilistic ways that schools have been and most likely will be chosen for evaluations of education reforms. Furthermore, the standard errors of fixed-effect estimators tend to be smaller than those for random-effect estimators because they need not account for how true impacts vary across schools. Thus, the statistical power of fixed-effect estimators is generally greater than that of random-effect estimators.[18]

However, random-effect estimators address more policy-relevant questions. For example, they focus on questions such as; "What is the average impact of an education reform for schools that might adopt it?"[19] In contrast, the corresponding fixed-effect question is: "What is the average impact of an education reform for schools in a particular study?" Because the first question is more general than the second it is relevant to a broader range of future decisions and decision-makers.

Indeed, it can be argued that no good evaluation should be relevant only to the sample studied because its results should useful for decisions about other situations. Thus, evaluators are often faced with a dilemma of choosing between: (1) fixed-effect estimators that are consistent with how sites were chosen but address questions which may be too narrow for many decision-makers, versus (2) random-effect estimators that are inconsistent with how sites were chosen but address broader policy questions.

When confronting this dilemma it is important to distinguish between two fundamentally different ways of making a generalization—s*tatistically* or *heuristically*. Statistical generalizations from samples of actual sites to populations of potential sites—like random-effect estimators—connote a high degree of specificity and rigor. This gives them a certain cachet that increases their weight as a source of scientific evidence. However, such weight is only warranted when the sample being used meets the statistical conditions required for the inferences being made. If these conditions are not met—which is almost always the case for evaluations of school reforms—then it is misleading to present findings that assume they do.

[18] In principle, it is possible that the standard error of a fixed-effect estimator (which gives equal weight to each school in a sample) could be *larger* than that of a random-effects estimator (which weights each school in proportion to the precision of its impact estimate) even though the latter estimator entails an additional source of random error (due to variation in true impacts across schools).

[19] For simplicity, the discussion in this section focuses only on average impacts for a group of schools. In principle, it could be extended to include the variance and other features of the distribution of impacts across schools.

In these situations, it might be preferable to generalize from a fixed-effect estimator in two separate steps. The first step would comprise a narrow but valid statistical inference to the sample of schools being studied. This step would address a question that is important in its own right: "What was the average impact for the schools studied?" Furthermore, it would address this question in a way that is consistent with how the schools were selected.

The next step would comprise a judgmental or heuristic attempt to generalize the study's findings to a broader range of potential situations. The breadth of this generalization should reflect the nature of the schools in the sample, why they chose to adopt the reform being studied, how they were chosen for the study, how they implemented the reform, and the conditions under which they implemented it. Evaluators should discuss these factors explicitly in a way that supports the generalization they believe is justified, specifies the likely limits to this generalization, and clarifies the nature of the information being used for it.

In practice, this two-step "heuristical" inference based on fixed-effect estimators might often be preferable to a one-step statistical inference based on random-effect estimators, because (1) it is consistent with how sites were selected and thereby provides valid answers to the questions it addresses, and (2) it is more explicit—and thus clearer—about what is known and what is not known about the generalizability of findings being presented. Therefore, it does not convey a spurious sense of scientific rigor.[20]

[20] The preceding argument is not meant to deny the importance of trying to estimate the true variation in impacts across sites when there are enough sites to do so with reasonable precision. If presented separately, such information can help to inform discussions about the generalizability of a study's findings, regardless of how its sites were chosen. The problem with random-effect estimators discussed above arises when cross-site variation is not addressed separately but is instead folded into the standard errors of combined impact estimates.

3. Estimating Impacts on Average Student Performance

This section describes how the preceding approach was used to measure the impacts of Accelerated Schools on the reading and math performance of third-grade students. The discussion outlines the analytic choices made, notes the considerations that went into these decisions and describes the final version of the approach selected. Impact findings from the evaluation are presented in Bloom et al. (2001).

3.1 Setting for the Analysis

MDRC's evaluation of Accelerated Schools examined standardized reading and math scores for third-grade students from eight elementary schools located in seven states. These schools had adopted an early version of the Accelerated Schools model during the mid-1990s.

The study focused on schools that reportedly had established "mature" programs. From the group of schools that were judged by the model's developers to have reached this stage, the evaluators chose a sample of schools that: (1) agreed to participate in the study, (2) had high concentrations of students at academic risk, (3) represented a broad range of urban settings, (4) had data available for individual third-grade reading and math scores from the same standardized test for at least five baseline years and five follow-up years, and (5) did not experience a major disruption or implement another education reform during the ten-year analysis period.

3.2 Overview of the Analysis

The impacts of Accelerated Schools on average third-grade reading and math scores were estimated separately for each school in the sample by comparing mean scores for each of five follow-up years to the school's *overall mean score for its three most recent baseline years*. A regression model was used to adjust for changes over time in student characteristics (which were minimal) and the variables included in these regressions depended on the student background data that were available for each school.[21] Cohort effects—which were small for reading and somewhat larger for math—were estimated using a variance component model, and the standard errors of all impact estimates were adjusted accordingly.

To combine findings across schools, test data for each were first transformed to Z-scores based on the mean and standard deviation for their three most recent baseline years. School-specific impacts, expressed as effect sizes, were then estimated from these Z-scores. Effect-size estimates were averaged across schools for each follow-up year, weighting each school equally. The standard errors for these average impacts were computed accordingly.

[21] Student characteristics available for most schools included gender, race or ethnicity, an indicator of over-age for grade, and an indicator for receipt of subsidized school meals.

The next two sections explain why the three-year baseline mean projection model was chosen and how cohort effects were estimated and used in the analysis.

3.3 The Baseline Projection Model

The first step in the selection of a baseline projection model was to plot the year-to-year pattern of mean baseline test scores for each school separately and for all schools combined. These plots were constructed using the regression adjusted Z-score for each sample member.[22]

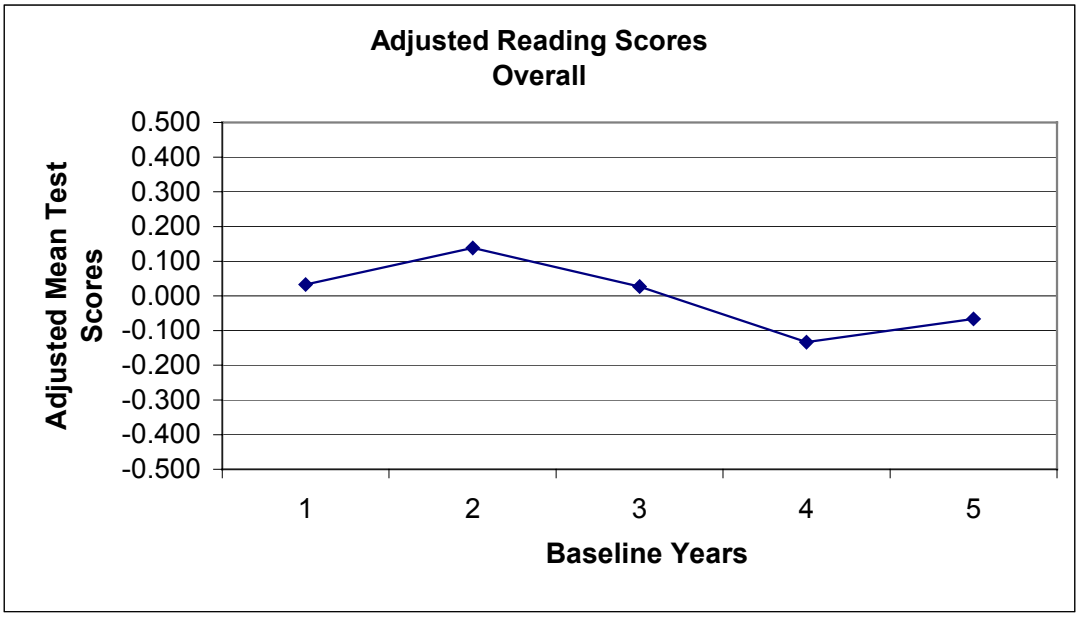
Figure 2 presents these graphs for all schools combined. The top panel in the figure is for reading and the bottom panel is for math. Points in the graphs are the regression-adjusted mean for each year measured in standard deviations above or below the baseline mean. As can be seen, these scores varied somewhat from year to year but did not exhibit a clear systematic upward or downward trend.

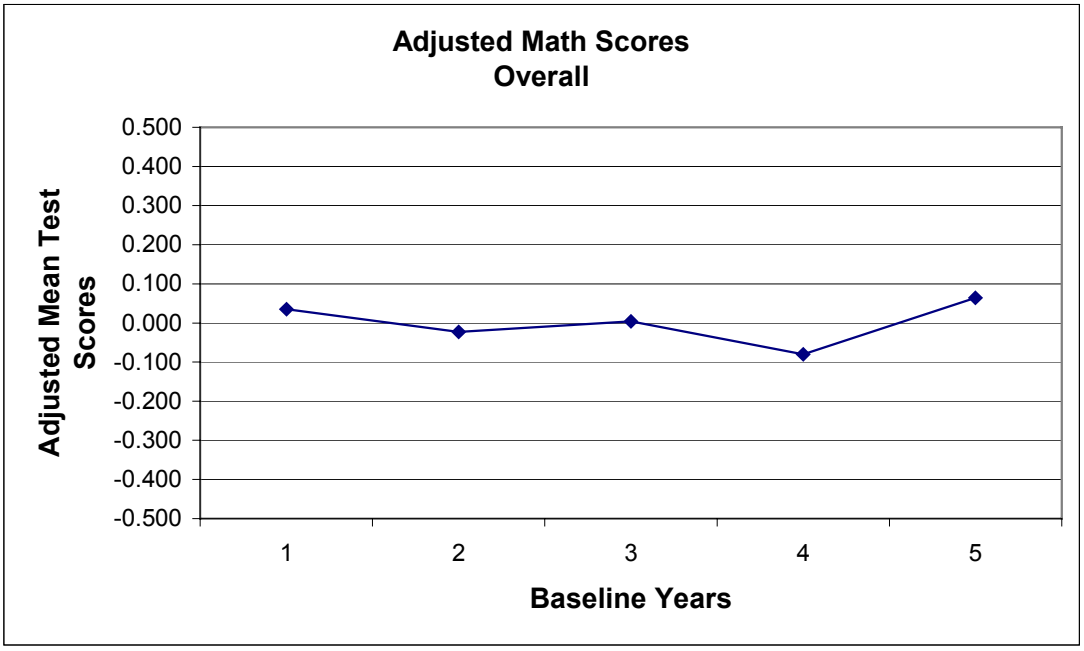
Baseline graphs for individual schools were inspected for systematic trends that might be masked by the combined findings in Figure 2. These graphs illustrated that individual schools had considerably more year-to-year fluctuation in mean scores than was the case for the combined findings. This is due to the smaller sample sizes for individual schools and their greater susceptibility to local, idiosyncratic events. Several schools exhibited an unusually high or low mean score (an outlier) for an early baseline year that distorted their baseline patterns. In addition, several schools exhibited a sharp upward or downward shift in their baseline scores. *But no school exhibited a continuous baseline trend that could be expected to continue for a number of years into the follow-up period.*

On balance then, visual inspection of the baseline performance of students in the Accelerated Schools sample indicated that a baseline trend model was not appropriate for estimating the impacts of the initiative.

[22] Because, as noted above, there were few changes over time in student characteristics, there were few differences between plots of regression-adjusted mean baseline test scores and their unadjusted counterparts, especially for all schools combined.

Figure 2
Regression-Adjusted Mean Annual Baseline Scores
for the Pooled Sample of Accelerated Schools





Corresponding evidence from a statistical analysis of baseline test data supports the same conclusion. This evidence was obtained by fitting a regression-adjusted linear trend model with cohort effects to the baseline Z-scores for students from each school. Estimated baseline slopes were then averaged across schools, weighting them equally.

Table 2 summarizes the results of this analysis. Findings for reading are presented in the top panel of the table and findings for math are presented in the bottom panel. The first column in the table lists the estimated baseline slope for each school separately and for all schools combined. The second column presents the estimated standard error for each slope and the third column indicates its statistical significance.[23]

Table 2

Estimated Baseline Slopes for the Accelerated Schools Sample

School	Baseline Slope	Standard Error of the Baseline Slope	Statistical Significance of the Baseline Slope[1]
		Reading	
School 1	-.19	.13	.15
School 2	-.10	.07	.17
School 3	.03	.04	.39
School 4	.02	.05	.68
School 5	.05	.05	.37
School 6	-.10	.05	.03
School 7	.01	.12	.91
School 8	-.07	.04	.09
Average	**-.04**	**.03**	**.12**
		Math	
School 1	.06	.05	.16
School 2	-.11	.07	.13
School 3	.06	.07	.40
School 4	-.01	.04	.87
School 5	.06	.10	.56
School 6	-.22	.07	.00
School 7	.08	.12	.51
School 8	.07	.05	.21
Average	**- .00**	**.03**	**.97**

[1] The p-value for a two-tailed t-test.

[23] The statistical significance level reported is the p-value for a two-tailed test of the null hypothesis that the slope was zero.

Note that the signs and magnitudes of the estimated baseline slopes vary widely across individual schools and subjects. Even for the same school the slope for reading can differ markedly from that for math. Given the considerable year-to-year fluctuation in mean test scores at a single school, only two out of sixteen baseline slopes were statistically significantly different from zero at the conventional 0.05 level, and they were for the same school. Only three slopes were statistically significant at the less stringent 0.10 level. *Furthermore, the slopes with the largest magnitudes reflected either an abrupt shift in baseline scores or a baseline outlier. They did not reflect a systematic trend.*

Now consider the combined slopes for all schools. Point estimates of these slopes suggest that mean baseline reading scores declined, on average, by 0.04 standard deviations per year during the baseline period, while those for math did not change overall. In addition, as illustrated by Figure 2, these slope estimates were not based on a clear pattern of change over time. Instead they reflect modest, year-to-year fluctuations. Therefore, neither the combined reading slope nor the combined math slope differs statistically significantly from zero.

Given the absence of empirical support for a baseline trend model and the substantial risks involved in projecting from a potentially incorrect slope (discussed in Section 2.1), a baseline mean model was used to project counterfactuals for the Accelerated Schools sample.

Having made this choice, it was then necessary to decide how many baseline years to use for projections. This question arose because of the shifts in the mean baseline scores of several schools and outliers that occurred during the earliest baseline years for several others. These *discontinuities* suggested that more recent baseline years might provide a more valid basis for projecting counterfactuals than would less recent (or all) baseline years. Thus, a decision was made to project counterfactuals from data for the three most recent baseline years.

3.4 The Variance Component Model of Cohort Effects

Having selected a three-year regression-adjusted baseline mean to project counterfactuals, it was then necessary to develop a corresponding procedure for estimating cohort effects. The options for doing so were limited, however, by the fact that three baseline years provide only two annual degrees of freedom per school to separate the cohort variance component from the individual variance component. To overcome this limitation, baseline data were pooled across schools and a common cohort effect was estimated from the total of 16 annual degrees of freedom available to do so. This estimate was then used to inflate standard errors for each school separately and for all schools combined.

The common cohort effect was determined using SAS PROC MIXED to estimate the following model (separately for reading and math) from a pooled sample of data for all baseline students.

$$Y_i = \sum_m \pi_m S_{mi} + \sum_m \sum_j C_{mj} S_{mi} X_{ji} + v_{mt} + e_i \qquad \textbf{(20)}$$

where:

Y_i = the test score for student i,
S_{mi} = one if student i was from school m and zero otherwise,
X_{ji} = background characteristic j for student i,
v_{mt} = a random error term for baseline year t in school m, and
e_i = a random error term for student i.

From the resulting maximum likelihood estimates of the variances for v_t and e_i (τ^2 and σ^2), cohort effects (ρ) of 0.002 for reading and 0.030 for math were obtained.[24] From these estimated cohort effects and the sample mean cohort size (n) of approximately 72 students, a Cohort Effect Multiplier, $\sqrt{1+(n-1)\rho}$, was computed for each subject. The resulting multipliers of 1.08 for reading and 1.77 for math were used to inflate OLS standard errors for estimates of impacts on mean test scores.

3.5 Combining Findings Across Schools

The preceding steps were taken to estimate program impacts on average student performance for each school based on regression-adjusted Z-scores for their students. Thus, all impact estimates were expressed as effect sizes. A simple mean (a fixed-effect estimator) was then used to combine these effect sizes across schools and a combined standard error was computed accordingly (Equation 17). This was done using a spreadsheet program.

A fixed-effect estimator was selected to combine findings across schools because they did not comprise a probability sample from an identifiable population. Thus, all *statistical* inferences were confined to the sample of schools that were part of the evaluation. When considering the implications of the study's findings for other schools, the evaluators tried to convey, in words, a sense of its likely generalizability given the nature of the schools involved and the education settings they represented.

3.6 Robustness of the Analysis

To test the robustness of the impact analysis with respect to variations in the projection model used, results from the following four models were compared using data from the Accelerated Schools evaluation: (1) the three-year regression-adjusted baseline mean used for the study, (2) a three-year baseline mean without regression adjustments, (3) a five-year regression-adjusted baseline mean, and (4) a five-year regression-adjusted baseline trend. Although these models produced somewhat different findings for individual schools, their pooled findings for all schools presented in Table 3 are generally consistent. Hence, the analysis for Accelerated Schools was fairly robust with respect to potential variations in the baseline projection model.

---

[24] The statistical significance of the resulting estimate of τ^2 was 0.374 for reading and 0.027 for math. Hence, this estimate was statistically significant for math (at the 0.05-level) but not for reading. Nevertheless, for both subjects, the point estimates for τ^2provide the best existing information about the likely magnitude of cohort effects. Thus, each estimate was used to compute a value for ρ, which in turn was used to compute a Cohort Effect Multiplier.

Table 3

Alternative Pooled Estimates of the Impacts of Accelerated Schools on Average Third-Grade Student Test Scores

Baseline Projection Model	Follow-up Year				
	One	Two	Three	Four	Five[1]
	Reading Impacts[2] (effect size)				
Three-Year Baseline *Mean* With Covariates[3]	- 0.044 (0.442)	0.019 (0.738)	- 0.148 (0.012)	0.019 (0.742)	0.190 (0.002)
Three-Year Baseline *Mean* Without Covariates	- 0.093 (0.180)	- 0.029 (0.679)	- 0.234 (0.001)	- 0.066 (0.341)	0.136 (0.077)
Five-Year Baseline *Mean* With Covariates	- 0.099 (0.264)	- 0.029 (0.745)	- 0.217 (0.015)	- 0.043 (0.630)	0.118 (0.236)
Five-Year Baseline *Trend* With Covariates	0.021 (0.850)	0.133 (0.311)	- 0.015 (0.924)	0.201 (0.252)	0.472 (0.035)
	Math Impacts[2] (effect size)				
Three-Year Baseline *Mean* With Covariates[3]	- 0.043 (0.648)	0.095 (0.312)	- 0.086 (0.378)	0.067 (0.481)	0.235 (0.027)
Three-Year Baseline *Mean* Without Covariates	- 0.077 (0.408)	0.055 (0.551)	- 0.156 (0.096)	- 0.006 (0.945)	0.207 (0.047)
Five-Year Baseline *Mean* With Covariates	- 0.048 (0.543)	0.102 (0.199)	- 0.095 (0.233)	0.054 (0.494)	0.224 (0.009)
Five-Year Baseline *Trend* With Covariates	- 0.062 (0.545)	0.085 (0.484)	- 0.112 (0.433)	0.035 (0.833)	0.240 (0.234)

[1] Data for this year were only available for seven of the eight schools in the Accelerated Schools study.
[2] The corresponding statistical significance level (p-value) for a two-tailed test is listed in parentheses below each impact estimate.
[3] This model was chosen for the Accelerated Schools analysis.

First consider the findings for math in the bottom panel of the table. The main value in each cell (the one that is not in parentheses) represents a pooled estimate of the impact of Accelerated Schools on average third-grade student performance for the eight schools in the study. This estimate is measured as an effect size in units of standard deviations. Below each

estimate in parentheses is a measure of its statistical significance (p-value). For example, during the first follow-up year, impact estimates range across the four models from - 0.043 to - 0.077 standard deviations. All of these estimates are quite small in magnitude and none is statistically significant. For subsequent follow-up years, findings from the models are also quite similar, with one slight exception—those from the second model, which does not control for student background characteristics. Impact estimates from this model are slightly less positive (or more negative) than those from the other models. This is probably due to the racial transition that occurred in one of the sample schools, which increased the importance of controlling for race in the analysis. Nevertheless, on balance, the results from all four models tell the same story.

The findings for reading presented in the top panel of the table are more sensitive to the baseline projection model used, but even they tell the same basic story for all models. Once again, the second model, which does not control for individual student background characteristics (and thus does not control for the racial transition that occurred in one school), produces impact estimates that are slightly less positive (or more negative) than those from the other models.

More striking, however, is the fact that the baseline trend model produces impact estimates that are more positive (and less negative) than those of the other models—and the difference increases noticeably for later follow-up years. This pattern reflects the finding presented earlier in Table 2 that the average baseline slope for reading was - 0.04 standard deviations per year.[25] Thus, toward the end of the follow-up period, the model projects a counterfactual (an estimate of what student performance would have been without Accelerated Schools) that is appreciably lower than that projected by the other models. This, in turn, produces impact estimates that are appreciably higher. For example, the baseline trend model produces an impact estimate of 0.472 standard deviations for the last follow-up year while corresponding estimates for the other models range from 0.118 to 0.190 standard deviations.

Recall, however, that the - 0.04 average baseline slope for reading *was not statistically significant and did not represent a systematic downward trend.* Rather it represented abrupt shifts in baseline scores at a couple of schools in the sample. Thus, the estimated slope was judged not to represent a systematic decline in scores that could be expected to continue for another five years into the follow-up period. Hence, the three-year regression-adjusted baseline mean was selected as the basis for estimating the impacts of Accelerated Schools.

3.7 Precision of the Analysis

A final important feature of the present analysis is its precision or minimum detectable effect. Intuitively, a minimum detectable effect is the smallest effect that an analysis has a "good" chance of detecting, if the effect exists. The smaller the minimum detectable effect is the more precise the analysis is. Bloom (1995) illustrates that the minimum detectable effect of an impact estimator equals a simple multiple of its standard error and shows how the value of the multiple depends on: (1) the p-value used to test whether the estimate is statistically significant,

[25] This was not an issue for estimating impacts on student performance in math because its average baseline slope was virtually zero.

(2) the desired statistical power of the test and (3) whether it involves a one-sided or two-sided alternative hypothesis.

The present discussion presents minimum detectable effects given conventional values of 0.05 for statistical significance, 80 percent for statistical power and a two-sided alternative hypothesis. For these parameters the minimum detectable effect of an impact estimator equals 2.8 times its standard error (Bloom, 1995). Because the present paper reports impact estimates as effect sizes, their precision is reported as *minimum detectable effect sizes*. Thus, for example, a minimum detectable effect size of 0.25 standard deviations indicates that a particular analysis has an 80 percent chance of finding an impact that is statistically significant at the 0.05 level using a two-sided hypothesis test—if a true impact of this magnitude exists.

Table 4 lists estimates of the minimum detectable effect size for third-grade reading and math performance as a function of the number of schools in one's sample. These results are based on pooled estimates of standard errors obtained using a three-year regression-adjusted baseline mean to project counterfactuals for the eight schools in the Accelerated Schools study.[26] The resulting minimum detectable effect sizes for a single school are 0.57 standard deviations for reading and 0.80 standard deviations for math.[27] These represent fairly large effects based on the conventional standards discussed in Section 2.5 (set by Cohen, 1988, and validated by Lipsey, 1990). In addition, they are much larger than the impacts of roughly 0.15 to 0.25 standard deviations produced by the Tennessee class-size experiment (Finn and Achilles, 1999). *Thus, impact estimates for individual schools are not at all precise.*

Table 4

Minimum Detectable Effect Sizes for Pooled Estimates of Impacts on Average Third-Grade Student Test Scores

Number of Schools in Sample	Minimum Detectable Effect Size for Reading	Minimum Detectable Effect Size for Math
1	0.46	0.75
5	0.20	0.33
10	0.14	0.24
15	0.12	0.19
20	0.11	0.17
25	0.09	0.15
50	0.06	0.11

Note: Results in this table were obtained by from a regression-adjusted baseline mean model applied to data for the eight schools in the Accelerated Schools evaluation.

[26] Minimum detectable effects for the two other baseline mean models are similar to those in Table 4. However, those for the baseline trend model are much larger, especially for later follow-up years. See Bloom (1999) for a discussion of minimum detectable effects for baseline trend models.

[27] The estimated minimum detectable effect size is larger for math than for reading because, as noted in Section 3.4, the estimated cohort effect multiplier is larger for math (1.77 versus 1.08).

However, as the number of schools in a sample increases, minimum detectable effect sizes for pooled impact estimates decline accordingly.[28] Thus, for 10 schools, the minimum detectable effect sizes are only 0.14 standard deviations for reading and 0.24 standard deviations for math. These are relatively modest impacts and fall within the range of those produced by the Tennessee class-size experiment. Thus, for many applications, samples of roughly 10 schools using the methodology presented in this paper might provide adequate precision for estimating average program impacts on average student performance. This was the case for the Accelerated Schools evaluation, which used a sample of 8 schools and observed reading and math impacts of 0.19 and 0.24 standard deviations—both of which were statistically significant.

[28] The minimum detectable effect for a pooled sample of n schools equals the minimum detectable effect for a single school divided by $\sqrt{n}$.

4. Estimating Impacts on the Distribution of Student Performance

To this point, the discussion has focused on measuring the impacts of school reforms on average student performance. However, it is possible—indeed likely—that different types of students will respond differently to reform initiatives and hence, the impacts they experience will vary. If so, then focusing only on average performance will mask variation that could provide insights about ways to improve these initiatives for important subgroups of students.

Thus, one should explore possibilities for learning about the overall *distribution of program impacts*. Unfortunately, this information cannot be obtained directly, because it is not possible to measure impacts for each sample member (by comparing his or her outcomes with and without the initiative being evaluated). Nevertheless, it is possible to estimate the impact of an initiative on the *distribution* of student performance and then speculate about where in the distribution impacts might have occurred. This approach produced important findings for the Accelerated Schools evaluation (Bloom et al., 2001).

The present section explains how this analysis was done. Specifically, it describes: (1) how the analysis was formulated, (2) how it was made operational, (3) how statistical significance was assessed, (4) how cohort effects were accounted for and (5) how impact estimates were combined across schools.

4.1 Formulating the Analysis

The distributional impact analysis for Accelerated Schools was formulated in a way that was consistent with the analysis of its impacts on average student performance. Both analyses compared student performance during each follow-up year for a school with a corresponding performance measure for its three most recent baseline years. Results for all schools were then averaged, weighting each equally. This process was carried-out separately for reading and math.

The first step in the process was to summarize the distribution of student performance during the baseline period. This summary was specified in terms of three baseline performance categories. An upper baseline category was defined as the top quartile or top 25 percent of all baseline scores. A middle baseline category was defined as the middle two baseline quartiles or middle 50 percent of all baseline scores. A lower baseline category was defined as the bottom quartile or 25 percent of all baseline scores. Hence, by definition, 25 percent of the baseline scores were in the upper baseline category, 50 percent were in the middle category and 25 percent were in the lower category.

Using the cut-off score for each baseline category it was possible to construct the distribution of scores across these categories for each follow-up year. This distribution was then compared with the baseline distribution to measure reform-induced change. Both the size and the statistical significance of this change were assessed.

Two further analytic steps were necessary in order to make this analysis comparable to that for average student performance: (1) regression adjustments for student characteristics, and (2) variance component estimates of cohort effects.

4.2 Making the Analysis Operational

The first step in the distributional analysis for a school was to estimate the following regression model of student test scores as a function of their background characteristics from pooled data for the three baseline years and five follow-up years.

$$Y_i = \alpha + \Sigma\delta_j X_{ji} + \nu_t + \varepsilon_i \qquad \textit{(21)}$$

The residuals from this model are regression-adjusted test scores. Thus, the upper baseline performance category was defined as the top 25 percent of the baseline residuals; the middle baseline category was defined as the middle 50 percent of these residuals; and the lower baseline category was defined as the bottom 25 percent. Performance for each follow-up year was then represented by the distribution of its residuals across the *baseline categories*.

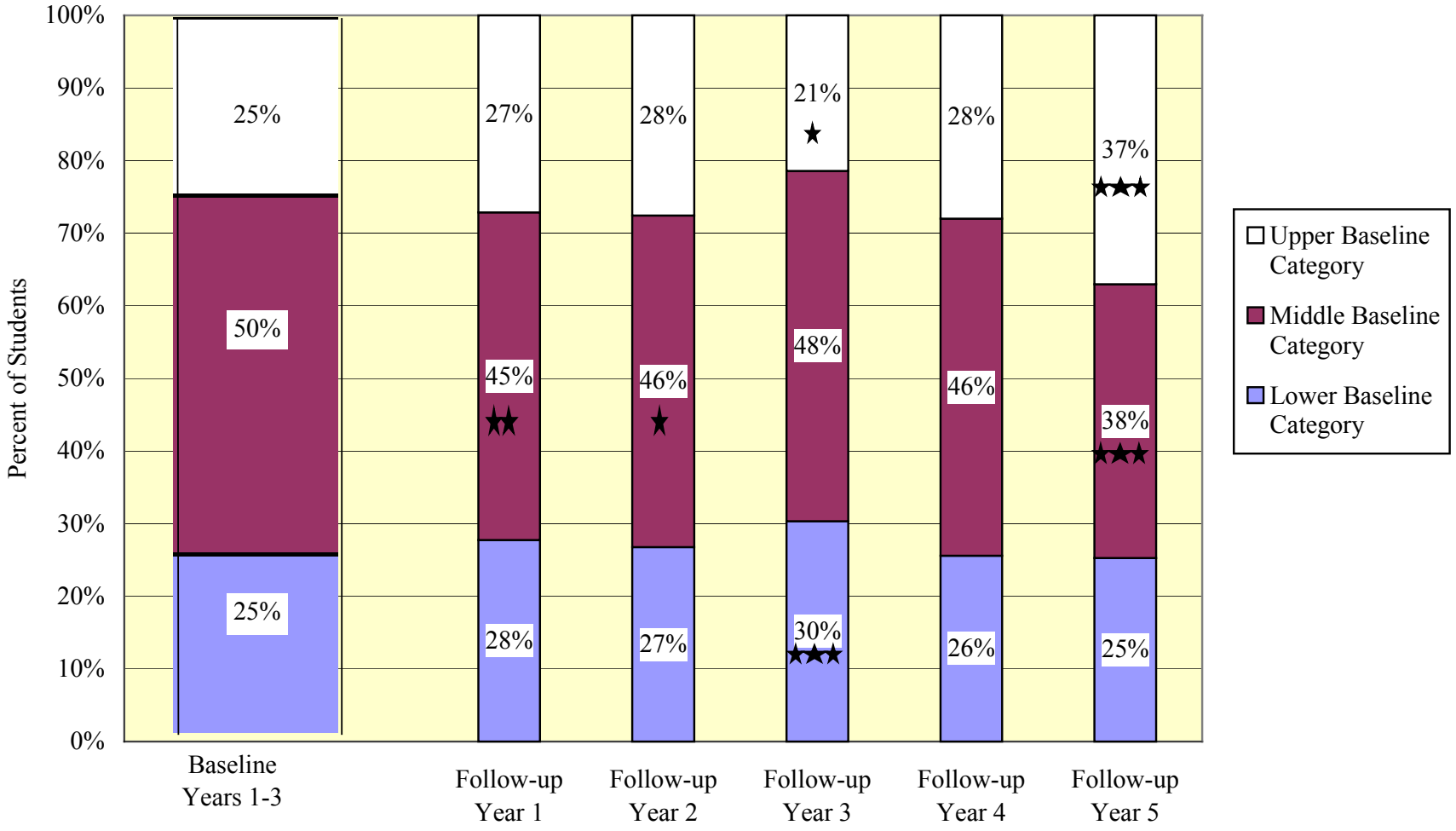
Consider, for example, the results displayed in Figure 3 for the average distribution of reading scores in the pooled sample of eight Accelerated Schools programs that were evaluated. The first (and widest) vertical bar in the figure represents the distribution of regression-adjusted reading scores (residuals) for third-grade students during the schools' *three* most recent baseline years. By definition, 25 percent of the baseline residuals are in the upper baseline category, 50 percent are in the middle baseline category and 25 percent are in the lower baseline category.

Note, however, that just because students score in the upper baseline category locally does not necessarily mean that they are above average statewide or nationally. Likewise, just because they score in the lower baseline category locally does not necessarily mean that they are below average statewide or nationally. Instead, their positions in their local distribution represent where they scored relative to third-graders from their school during its baseline period. Because many education reforms have focused on low-performing schools, it is likely that many students in the upper and upper middle portions of their local baseline distributions are below average statewide or nationally. Thus, it is important to keep this in mind when interpreting the findings of a distributional impact analysis.

Now consider how the distribution of regression-adjusted test scores (residuals) in the figure changed over time. For the first two follow-up years there was virtually no change. This might occur if the education reform being evaluated focused initially on changing institutional arrangements and school decision-making processes instead of improving instructional practices, as was the case for the early-vintage Accelerated Schools programs that were evaluated.

In subsequent years however, the distribution of regression-adjusted test scores (residuals) began to change. During the third follow-up year, scores declined overall, with an increase in the percentage that fell within the lower baseline category (to 30 percent) and a decrease in the percentage that fell within the upper category (to 21 percent). This might reflect confusion generated by initial incomplete attempts to change classroom practices. However, during the next two years, test scores began to rise. Thus, by the fifth and final year of the follow-up period, the average school in the study had 37 percent of its regression-adjusted scores in the upper baseline category. This was both substantially and statistically significantly

Figure 3
The Average Distribution of Baseline and Follow-up Reading Scores for the Eight Accelerated Schools



(discussed later) higher than the 25 percent of scores that were in this category during the baseline period.

However, there was virtually no change in the percentage of scores in the lower baseline category. Thus, it appears that the reform did not affect the performance of the weakest students at the schools in the study. It only appears to have moved students who otherwise would have scored in the middle baseline category (who were about average locally) into the upper category (which was above average *locally)*.[29]

As noted above however, the impact of an education reform on the distribution of student performance cannot identify *with certainty* the distribution of impacts on performance. Thus, it cannot identify with certainty which students experienced what impacts. This is because there are many possible explanations for any observed pattern of impacts on the performance distribution.

For example, in Figure 3 it is possible that some students who would have scored in the lower category without the reform experienced positive impacts that were large enough to move them into the upper category, while an equal number of students who would have scored in the middle category experienced negative impacts that moved them into the lower category. Although this alternative explanation is logically possible, and thus cannot be ruled-out with certainty, it is highly implausible. Instead the most plausible interpretation of the particular pattern of impacts illustrated by the figure is that students who were about average locally improved their performance. There is little evidence that the reform improved the performance of students who were below average locally.[30] One possible explanation for this result is the fact that weaker students tend to move more frequently, and thus may have had less exposure to the reform.

4.3 Assessing Statistical Significance

Before trying to interpret the magnitude and direction of observed changes in the distribution of test scores, one should assess their statistical significance. If these changes are not statistically significant then they might just represent random error.

There are two different ways to specify a test of the statistical significance of a difference between a baseline performance distribution and its counterpart for a given follow-up year. These specifications represent two different types of hypothesis tests: (1) "omnibus" tests and (2) "category-specific" tests. An omnibus test assesses the statistical significance of an observed change in the full distribution of student performance. It tests the null hypothesis of *no change for any category.* A category-specific test assesses the statistical significance of an observed change for a particular category. It tests the null hypothesis of *no change for that category*.

[29] These scores were still below average statewide or nationally in many cases.

[30] One could take this analysis one step further to see whether the average score in the lower baseline category was higher during the follow-up period than it was during the baseline period. This could help to determine whether students who were below average locally had been helped by the reform but not by enough to move them into a higher performance category.

An omnibus test of the difference between a baseline and follow-up distribution could be conducted from a cross-tabulation of their residuals by performance category. The chi-square statistic for this three-by-two cross-tabulation tests the null hypothesis of no real difference between the two performance distributions. Thus, it incorporates information about observed differences for all performance categories.

Unfortunately, it is not clear how to account for cohort effects in an omnibus test. Thus, if cohort effects exist (which findings presented later suggest is the case for math) the chi-square statistic will not reflect the full extent of random year-to-year variation in student performance and will overstate the statistical significance of observed changes. Therefore, if the test indicates that a change is significant, this finding is inconclusive. If, on the other hand, the test indicates that a change is not significant, it is conclusive.

A category-specific test of a follow-up versus baseline difference in the percentage of residuals in a *particular category* can be specified as a two-sample difference of proportions, with $\hat{P}_F$ for the follow-up year and $\hat{P}_B$ for the baseline period. Given the total sample size for the follow-up year (n_F) and the baseline period (n_B) plus a measure of the cohort effect multiplier (CEM) one can use the following t-statistic for this test.[31]

$$t = \frac{\hat{P}_F - \hat{P}_B}{SE(\hat{P}_F - \hat{P}_B)} \qquad \textbf{(22)}$$

where:

$$SE(\hat{P}_F - \hat{P}_B) = CEM\sqrt{\frac{\hat{P}(1-\hat{P})(n_F+n_B)}{n_F n_B}} \qquad \textbf{(23)}$$

and

$$\hat{P} = \frac{n_F \hat{P}_F + n_B \hat{P}_B}{n_F + n_B} \qquad \textbf{(24)}$$

[31] Because the baseline proportion for the category was established by definition, it might seem that the comparison should be specified as a one-sample test of the *estimated* follow-up proportion versus the *known* baseline proportion of 0.25. However, in order to reflect the random error that actually exists in the baseline distribution, it seems more appropriate to specify the comparison as a two-sample test. This implies that the cut-off used to define the baseline category was a convention of the analysis instead of a known quantity of interest.

4.4 Accounting for Cohort Effects

To implement the preceding category-specific t-test it is necessary to estimate the cohort effect and the corresponding Cohort Effect Multiplier. This was done for the Accelerated Schools evaluation using SAS PROC MIXED software to estimate the variance component structure of the following linear probability model, separately by performance category and subject.

$$P_i = \sum_m \gamma_m S_{mi} + \eta_{mt} + \varepsilon_i \qquad \textit{(25)}$$

where:

P_i = one if the residual for student i was in the baseline performance category of interest and zero otherwise,
S_{mi} = one if student i was from school m and zero otherwise,
η_{mt} = a random error term for baseline year t in school m, and
ε_i = a random error term for student i.

Equation 25 was estimated from pooled baseline data (three years per school) for the eight schools in the sample.

The estimated cohort effect, $\rho = \tau^2/(\tau^2+\sigma^2)$, was zero for the three performance categories in reading. This implies a Cohort Effect Multiplier of one for each category. Thus, OLS standard errors were used directly to assess the statistical significance of category-specific impacts on reading performance. For math, the estimated cohort effects were 0.028, 0.002 and 0.013 for the upper, middle and lower baseline performance categories, respectively. The corresponding Cohort Effect Multipliers of 1.74, 1.07 and 1.38[32] were used to inflate OLS standard errors for each category of math performance.[33]

4.5 Combining Findings Across Schools

Through a spreadsheet program, a fixed-effect estimator was used to combine distributional impact findings across the Accelerated Schools sample. The first step in this process was to combine percentage estimates across schools for each performance category during the baseline period and each follow-up year. The output of this step was a bar chart like Figure 3 for the combined sample of schools. The second step was to test the statistical significance of category-specific changes from the baseline period to each follow-up year. The output of this step was a pattern of stars on the bar chart indicating which categories in which years (if any) differed significantly from their baseline counterpart.

---

[32] All Cohort Effect Multipliers were based on a cohort size of 72 students, which was approximately the average annual grade size for the Accelerated Schools analysis sample.

[33] The statistical significance of τ^2 (and hence, the approximate significance of ρ) was 0.055, 0.715 and 0.186 for the upper, middle and lower math baseline performance categories, respectively. Thus, only the estimate for the upper category was statistically significant. Nevertheless, because these point estimates represent the best existing information about cohort effects, all three estimates were used to inflate OLS standard errors.

4.5.1 Combining Performance Distributions Across Schools

For any given period and baseline performance category, findings were combined across schools by taking the mean percentage of students in that category, weighting each school equally. However, to simplify notation, the following discussion is stated in terms of proportions instead of percentages. Thus, the first step for a given performance category and follow-up year was to compute its mean proportion for all schools, where:

$$\bar{P}_F = \sum_{1}^{m} \hat{P}_{Fj} / m \qquad \textit{(26a)}$$

and:

$\bar{P}_F$ = the mean proportion of students in the performance category during follow-up year F,

$\hat{P}_{Fj}$ = the proportion of students in the performance category at school j during follow-up year F, and

m = the total number of schools.

The mean proportion of students in each category during the baseline period was 0.25, 0.50 and 0.25 because these proportions were defined to be the same for all schools. However, to clarify the discussion of standard errors below, it is useful to note that they can be obtained as:

$$\bar{P}_B = \sum_{1}^{m} \hat{P}_{Bj} / m \qquad \textit{(26b)}$$

Having computed the mean proportion for each performance category for each follow-up year and the baseline period, these results were plotted as percentages in a bar chart like Figure 3.

4.5.2 Assessing the Statistical Significance of Category-Specific Changes

The next step was to assess the statistical significance of differences between the combined performance distribution for each follow-up year and that for the baseline period. This was accomplished by testing the significance of category-specific differences. Thus, for each performance category in each follow-up year the following t-test was conducted

$$t = \frac{\bar{P}_F - \bar{P}_B}{SE(\bar{P}_F - \bar{P}_B)} \qquad \textit{(27)}$$

$$SE(\bar{P}_F - \bar{P}_B) = CEM\sqrt{VAR(\bar{P}_F) + VAR(\bar{P}_B)} \qquad \textit{(28)}$$

$$VAR(\bar{P}_F) = \sum_{1}^{m} VAR(\hat{P}_{Fj}) / m \qquad \textbf{(29a)}$$

$$VAR(\bar{P}_B) = \sum_{1}^{m} VAR(\hat{P}_{Bj}) / m \qquad \textbf{(29b)}$$

$$VAR(\hat{P}_{Fj}) = \hat{P}_{Fj}(1 - \hat{P}_{Fj}) / n_{Fj} \qquad \textbf{(30a)}$$

$$VAR(\hat{P}_{Bj}) = \hat{P}_{Bj}(1 - \hat{P}_{Bj}) / n_{Bj} \qquad \textbf{(30b)}$$

The number of degrees of freedom for this test is approximately equal to the total number of students in the baseline and follow-up samples being compared *minus* the number of parameters (coefficients plus the intercept) estimated in the regression model used to create residual tests scores *minus* two.[34]

[34] Because this test was based on relatively large samples of students, its t-distribution closely approximates a z-distribution and its findings are not sensitive to the exact number of degrees of freedom involved.

5. Estimating Impacts on the Variation in Student Performance

This section presents a method for estimating impacts on the variation in student performance. The discussion first describes the analysis for a single school and then explains how to combine findings across schools.

5.1 The Issue

The variability of student performance in reading or math is a measure of disparity in basic academic proficiencies and for many children this early academic disparity will become a later economic disparity. Widespread concern about the linkage between these two unwanted outcomes has been a major factor motivating education reforms. Thus, when evaluating a reform it is important to examine its impacts on the variation in student performance.

For example, as noted above, the findings in Figure 3 suggest that the Accelerated Schools reform improved the performance of third-grade students who were in the middle of their local performance distribution without helping students at the lower end. This shift represents an increase in the variability of student performance and thus an increase its disparity. Therefore, the evaluation attempted to quantify this result and assess its statistical significance.

5.2 Measuring Within-Cohort Variation in Student Performance

The first step in an analysis of program impacts on the variation in student performance is to define the construct of interest. For this purpose one should focus on the variation in performance within annual cohorts of students in a particular grade and ask the question: "How did the education reform being evaluated affect this within-cohort variation?"[35]

There are two standard parameters for measuring the variation of any distribution—its variance and its standard deviation—and either one could be used to measure variation in student performance. The standard deviation was used for the evaluation of Accelerated Schools because it is defined in the same units as the test scores being summarized (points on a scale). In contrast, the variance is defined in terms of the square of these units (scale-points squared).

Nevertheless, the standard deviation of a test-score distribution is difficult to interpret in absolute terms. Instead, it is more useful as a comparative measure—either across schools (to assess differences) or over time (to gauge change). Thus, it provides a useful metric for comparing impacts but it is less useful for assessing their absolute magnitude.

5.3 Measuring Program Impacts on Within-Cohort Variation

The Accelerated Schools analysis of impacts on the variation in student performance was formulated in a way that was consistent with analyses of impacts on the average and distribution of performance. Impacts on performance variability were estimated by comparing the standard

[35] This question should be addressed for all grades that have available data.

deviation of regression-adjusted test scores (residuals) for each annual follow-up cohort to the pooled within-cohort standard deviation for the three most recent baseline years.

The parameter used to compare the standard deviation for a given follow-up cohort, S_F, with its baseline counterpart, S_B, was the ratio of the two or $R_{FB} = S_F/S_B$.[36] For example, a ratio of 1.07 for the first follow-up year for a given school, would indicate that the standard deviation for that cohort was 7 percent larger than that for the baseline cohort.

5.4 Assessing the Statistical Significance of Impact Estimates

The next step is to assess the statistical significance of the observed difference between the standard deviation for a follow-up cohort, S_F, and that for the baseline period, S_B. The null hypothesis of *no difference* between these two measures implies no difference between their variance counterparts, S_F^2 and S_B^2. This in turn, implies a ratio of one for the standard deviation and a ratio of one for the variances.

The alternative hypothesis of *some difference* between the two standard deviations has corresponding implications for their variance and ratio counterparts. If the standard deviations are significantly different from each other, then so are their variance counterparts. In addition, their ratio counterparts will differ significantly from one.

Furthermore, the alternative hypothesis of some difference between the standard deviation for a specific follow-up cohort and that for the baseline period leaves open the possibility that either one could be larger than the other. This is consistent with the possibility than an education reform could reduce or increase the variation in student performance. Hence, the alternative hypothesis of interest is two-sided, which has important implications for the statistical test.

The simplest way to specify this test is through the variance ratio, S_F^2/S_B^2, which has a standard F distribution with approximately $(n_F\text{-}c\text{-}1)$ degrees of freedom for the numerator and $(n_B\text{-}c\text{-}3)$ degrees of freedom for the denominator. The c term is the number of regression coefficients (including the intercept) in the model used to create regression-adjusted test scores (residuals).

One can use an F distribution to test a null hypothesis of no difference between two standard deviations (through their variance counterparts) against a two-sided alternative hypothesis of some difference as follows. First, always put the larger variance in the numerator. Then, to account for this fact, double the p-value for the resulting F-statistic.

[36] The pooled baseline standard deviation was defined as the square root of the mean of the three annual within-cohort variances. The analysis described in this section will produce identical results for regression-adjusted scale scores or regression adjusted Z-scores.

5.5 Combining Impact Estimates Across Schools

To combine findings across the Accelerated schools sample, the mean variance for each follow-up year and its baseline counterpart were computed, weighting each school equally. The ratios of the corresponding standard deviations were used to summarize how the overall average variability in student performance changed over time. For reading, these ratios were 1.07, 1.07, 1.03, 1.05 and 1.13 for each follow-up year respectively.

The ratios of the combined variances were used as an F test of the significance of these observed changes. The number of degrees of freedom for the numerator and denominator of this test were set equal to their corresponding totals for all schools in the sample. Based on this test the first two ratios of 1.07 were statistically significant at the 0.10 level, the last ratio of 1.13 was significant at well beyond the 0.01 level, and the other two ratios of 1.03 and 1.05 were not statistically significant.

6. Toward a More Comprehensive Quasi-Experimental Strategy

Methodological work is currently underway at MDRC to extend the interrupted time-series methodology described in this paper. In addition, applications of these extensions are planned for major evaluation studies of Project GRAD (Ham, Doolittle and Holton, 2000), First Things First (MDRC, 2000a) and Talent Development Schools (MDRC, 2000b).

Three main extensions of the approach are being considered: (1) the addition of comparison series, (2) the addition of value-added analysis, and (3) the addition of hierarchical growth path models. Each of these extensions brings with it new strengths and weaknesses. Thus, it is hoped that when used together, the strengths of each will help to offset the weaknesses of the others. Furthermore, none of the extensions is feasible in all situations. Therefore, it is hoped that by creatively and carefully using different combinations for different situations, one can provide rigorous answers to the evaluation questions of greatest interest for the broadest possible range of settings and circumstances.

6.1 Adding Comparison Series

A comparison series design adds an interrupted time-series analysis for comparison schools that did not adopt an education reform but maintained time-series data on student performance that was comparable to that for schools that did adopt the reform (program schools). The comparison schools' deviation from their baseline performance pattern provides a new estimate of the counterfactual for program schools.

This estimate helps to control for other local changes that might have affected student performance while the reform was being implemented at program schools. Thus, it helps to protect the internal validity of program impact estimates against the methodological threat of "local history."[37] Intuitively the comparison school deviation from its baseline performance pattern indicates what the program school deviation would have been without the reform. Bloom (1999) describes how to use a comparison series design and examines its statistical power.

Obtaining data for comparison schools (which often have nothing to gain from participating in an evaluation and much to lose from potentially invidious comparisons to program schools) can be much more difficult than obtaining data for program schools (that want to demonstrate success). This can be especially problematic for evaluations of reforms being tested in multiple school districts where cooperation is required from many different organizations and decision-makers.

Thus, in practice, a comparison series design might only be feasible for evaluations conducted in a small number of districts. In this case, it might be possible to obtain a time-series

[37] The main source of protection against "local history" for the Accelerated Schools evaluation was *replication.* Its sample of eight schools from eight school districts and seven states provide eight independent tests of the reform model. Thus, while idiosyncratic local events might undermine the causal interpretation of findings for one school, they provide a less plausible alternative explanation for a systematic pattern of change for a group of schools. The more schools in the sample and the more independently of each other they operate, the more methodological protection replication provides.

of performance data for program and comparison schools from the same central source. Or if not, district leaders might help researchers acquire time-series data from comparison schools within their domain.

6.2 Adding Value-added Analysis

Value-added analysis, which was discussed briefly in Section 2.2, measures student achievement as a function of education inputs. In its simplest form, value-added analysis represents a posttest-pretest design with a comparison group. It estimates program impacts on posttest scores by controlling statistically for pretest scores and student background characteristics.[38]

For example, one might compare eighth-grade math achievement scores for program schools with those for comparison schools, controlling statistically for each student's sixth-grade scores and background characteristics.[39] Doing so would reflect the difference between program schools and comparison schools in their *increments* to math achievement—their value added—between grades 6 and 8.

Although highly appealing in many ways, and widely used for education research, evaluation and performance measurement, valued-added analysis has an important limitation due to potential selection bias. This bias will exist if there are underlying differences between student *growth paths* at the program schools and comparison schools.

In the same way that it is possible for the underlying performance level of program and comparison schools to be different because of unmeasured student characteristics, it is also possible for their students to have different underlying growth paths. Thus, observed differences between their student performance gains might reflect both the underlying differences in their growth paths plus any differences that were produced by the education reform being evaluated.

If student growth paths are inherently steeper, on average, at program schools than at comparison schools, then value-added estimators will tend to overstate program impacts. On the other hand, if student growth paths are inherently less steep at program schools, then value-added estimators will tend to understate program impacts.

To deal with this potentially important weakness of value-added analysis while simultaneously benefiting from its considerable strengths, it is possible in some settings to combine the approach with interrupted time-series analysis. To do so requires data on posttests

---

[38] As noted earlier, the most important feature of value-added analysis is controlling for pretest scores. Controlling for student background characteristics may not add much beyond this.

[39] 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 eighth-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. Another common way to formulate the model is to specify the dependent variable as the difference between each student's posttest and pretest (their gain score). This is equivalent to setting b_1 equal to one.

and pretests for a number of baseline cohorts before a school reform was launched and a number of follow-up cohorts after the reform was launched. For each cohort, one could estimate the mean gain in performance from students' pretest to their posttest. This would produce a time-series of performance *gains* instead of performance *levels* (the focus of the present paper). One could then seek to determine how, if at all, performance gains during the follow-up period deviated from the pattern of performance gains during the baseline period.

The focus on performance gains reflects the value-added component of the combined design and the focus on follow-up deviations from baseline patterns reflects its interrupted time-series component. By controlling for observed changes over time in cohort pretest scores, the value-added component can help to protect the interrupted time-series component against bias from changes over time in student abilities and backgrounds.

To further strengthen the design, one could add a comparison series to it. This would require comparable time-series data on pretests and posttests for cohorts of students at the comparison schools. From these data one could conduct an interrupted time-series analysis of the baseline to follow-up deviation in performance gains at the comparison schools. Program impacts on average student value-added could then be obtained from the difference between deviations for the program and control schools.

This three-way combination design (interrupted time-series analysis *plus* value added analysis *plus* a comparison series analysis) deals directly with the primary weaknesses of its principle components. The comparison series analysis helps to control for local idiosyncratic events that might change student performance, which is a potential problem for interrupted time-series analysis. The value-added analysis helps to control for shifts in the composition of student cohorts over time, which is another potential problem for interrupted time-series analysis. The interrupted time-series analysis with a comparison series helps to control for underlying differences in student growth paths at program and comparison schools, which is a potential problem for value-added analysis.

6.3 Adding Hierarchical Models of Student Growth Paths

One weakness of value-added analyses that are formulated as single pretest-posttest comparisons grows out of the long-standing controversy in the social sciences over how to analyze change data.[40]

On the one hand, change in student performance could be analyzed through gain scores obtained from each student's pretest to posttest difference in scores. This assumes, either explicitly or implicitly, a particular model of the relationship between posttests and pretests. Specifically, it assumes that on average, a student's posttest equals his or her pretest plus a constant.

On the other hand, one could examine posttests directly using a regression model to control statistically for pretests. This assumes a less restrictive model of the relationship between

[40] See Oakes and Feldman (2001) for a recent discussion of this controversy.

pretests and posttests, but it tends to under-adjust for preexisting differences among students because the pretest coefficient tends to be underestimated.[41]

To address this problem and thereby take the proposed quasi-experimental strategy one step further, it might be possible to expand its value-added component from a single pretest and posttest to multiple pretests and posttests per student. One could then use a two-level hierarchical model to estimate the growth path for each student and account for annual cohort effects at each school.[42]

To estimate individual student growth paths requires at least three, and preferably four or more annual tests per student. For example, if the reading performance of elementary school students were tested in grades 2, 4, 5, and 6, one could use these scores to estimate a linear growth path for each student.[43] However, this would only be meaningful if the tests produced a comparable measure of the same construct across grades in a cumulative cardinal metric.[44]

Under these conditions, which exist at some schools and in some school districts, one could formulate a value-added analysis in terms of the *slopes* of students' growth paths.[45] Doing so would express value-added as the average annual increase in student performance.[46] One could also measure the performance *level* for any given grade in terms of the average value on students' growth paths for that grade.

By imbedding this hierarchical value-added analysis within an interrupted time-series analysis that included a comparison series, one could maximize the methodological leverage of the overall quasi-experimental strategy. Intuitively, the analysis would proceed as follows:

It would begin by computing the slope of the growth path for each student. This would measure students' average annual increase in performance (their average annual value-added).[47] For program schools, one could then compute the average slope of the growth paths for students in each baseline and follow-up cohort (with or without statistical controls for individual characteristics).

The interrupted time-series component of the design would then use the baseline *pattern* of mean growth path slopes to project a counterfactual for the follow-up cohorts. The difference

---

[41] This problem of attenuation due to measurement error occurs in many settings (Greene, 1997, p. 437). The problem does not exist for gain scores, however, because they constrain the regression coefficient for pretests to equal one.

[42] It is possible to use a three-level hierarchical linear model to also account for how annual cohorts are grouped within schools. This would specify the variation in impacts across schools as random effects. However, given this paper's emphasis on fixed-effect estimators to combine findings across schools, the third level of analysis for schools would be addressed outside of the hierarchical model.

[43] With four tests per student it is conceptually possible but probably not operationally practical to estimate nonlinear growth paths as well.

[44] Bryk et al. (1998) discuss this issue in detail.

[45] This approach has been used by Ross, Sanders and Wright (2000) and by Bryk et al. (1998), and it is growing in popularity.

[46] Variants of this valued-added approach also control for certain characteristics of students and schools (see Raudenbush and Willms, 1995, and Meyer, 1997)

[47] The slope of the growth path in the hierarchical analysis replaces the gain score in the pretest/posttest analysis.

between the actual and projected mean slopes for each follow-up cohort would thus, provide an estimate of the impact of a school reform on value-added

This analysis could then be repeated for a group of comparison schools. Doing so would produce estimates of the deviation of their follow-up growth path slopes from their baseline counterparts. This deviation for comparison schools could then be used to infer what the deviation would have been for program schools without the reform. Hence, the impact of the reform on student-value added could be obtained from the difference between the two deviations.

6.4 Tailoring the Evaluation Design to the Education Setting

A complete version of the preceding quasi-experimental strategy requires high quality data for large samples of students and long periods of time. This means that only rarely will it be possible to use the complete strategy in the near future. Thus, in the short run, while it is necessary to take existing testing procedures and data systems as given, the key to effective use of the strategy is an opportunistic effort to apply as many of its components as possible in any given setting. To help support these applications it will be necessary to work through the methodological implications of different combinations of components and to establish effective procedures for implementing them.

However, as more is learned about the strategy—both in terms of the costs of applying it and the benefits of doing so—there may be advances in testing methods and data systems that facilitate greater use of more expanded versions of the analysis. Indeed, as the value of the approach becomes better understood, performance measurement systems could be developed explicitly to support its use.

Nevertheless, both in the near term and in the longer term, because of the decentralized way that primary and secondary education are provided in the United States, it will remain necessary to carefully tailor each application of the strategy to meet the needs and capabilities of the local setting in which it is used. In short, there will never be a time when “one size fits all.”

References

Bloom, Howard S. 1999. "Estimating Program Impacts on Student Achievement Using 'Short' Interrupted Time-Series." New York: MDRC.

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review*, 19(5): 547-556.

Bloom, Howard S., Sandra Ham, Laura Melton, and Julieanne O'Brien. 2001. *Evaluating the Accelerated Schools Approach: A Look at Early Implementation and Impacts in Eight Elementary Schools*. New York: MDRC.

Bryk, Anthony S., and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, Calif.: Sage Publications.

Bryk, Anthony, Yeow Meng Thum, John Q. Easton, and Stuart Luppescu. 1998. *Academic Productivity of Chicago Public Elementary Schools.* Chicago: Consortium on Chicago School Research.

Campbell, Donald T., and Julian C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd edition). Hillsdale, N.J.: Lawrence Erlbaum.

Cook, Thomas, and Donald T. Campbell. 1979. *Quasi-Experimental Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Cooper, Harris, and Larry V. Hedges (eds.). 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Finn, Jeremy D., and Charles M. Achilles. 1999. "Tennessee's Class Size Study: Findings, Implications, Misconceptions." *Educational Evaluation and Policy Analysis*, 21 (2): 97-109.

Greene, William H. 1997. *Econometric Analysis*. Upper Saddle River, N.J.: Prentice Hall.

Ham, Sandra, Fred C. Doolittle, and Glee Ivory Holton. 2000. *Building the Foundation for Improved Student Performance: The Pre-Curricular Phase of Project GRAD Newark.* New York: MDRC.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Boston: Academic Press, Inc.

Lipsey, Mark. 1990. *Design Sensitivity: Statistical Power for Experimental Research* (pp. 51-56). Newbury Park, Calif.: Sage Publications.

MDRC. 2000a. *The Evaluation of First Things First Research Design Report.* New York: MDRC.

MDRC. 2000b. *An Evaluation of the Talent Development Model: Research Design Report.* New York: MDRC.

Meyer, Robert H. 1997. "Value-Added Indicators of School Performance: A Primer." *Economics of Education Review*, 16(3): 283-301.

Murray, David M., and Brian Short. 1995. "Intra-Class Correlations Among Measures Related to Alcohol Use by Young Adults: Estimates, Correlates and Applications in Intervention Studies." *Journal of Studies on Alcohol*, 56(6): 681-693.

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*, 21 (2): 127-142.

Oakes, J. Michael, and Henry A. Feldman. 2001. "Statistical Power for Nonequivalent Pretest-Posttest Designs." *Evaluation Review*, 25 (1): 3-28.

Olejnik, Stephen, and James Algina. 2000. "Measures of Effect Size for Comparative Studies: Applications, Interpretations and Limitations." *Contemporary Educational Psychology*, 25: 241-286.

Raudenbush, Stephen W., and J. Douglas Willms.1995. "The Estimation of School Effects." *Journal of Educational and Behavioral Statistics*, 20 (4), 307-335.

Sanders, William L., and Sandra P. Horn. 1994. "The Tennessee Value-Added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment." *Journal of Personnel Evaluation in Education*, 8: 299-311.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. Forthcoming. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin.