MDRC Working Papers on Research Methodology

Estimating Program Impacts on Student Achievement Using "Short" Interrupted Time Series

Howard S. Bloom

Manpower Demonstration
Research Corporation

MDRC

August 1999

This is part of a new series of papers that explore alternative methods of examining the implementation and impacts of programs and policies.

This paper was prepared for the 1999 Annual Meeting of the American Educational Research Association in Montreal, Canada. Please send questions or comments to:
howard_bloom@mdrc.org.

Dissemination of MDRC's work is supported by MDRC's Public Policy Outreach Funders: the Ford Foundation, the Ambrose Monell Foundation, the Alcoa Foundation, and the James Irvine Foundation.

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

For information about MDRC, see our Web site: www.mdrc.org.
MDRC® is a registered trademark of the Manpower Demonstration Research Corporation.

ABSTRACT

This paper explores the use of interrupted time-series analysis for estimating the impacts of school restructuring programs designed to increase student achievement. The paper first illustrates the approach and considers its strengths and weaknesses. It then describes how to estimate program impacts and their standard errors from a simple regression model. Next it focuses on the statistical precision of these impact estimates (their minimum detectable effect size) and the research design considerations that affect this precision. The paper concludes by briefly outlining several important issues related to implementing the approach that will be addressed in future research.

This paper explores the use of interrupted time-series analysis for estimating the impacts of programs designed to increase the academic achievement of students in primary and secondary school. This method is particularly relevant for evaluating programs designed to produce school-wide change.

The goals of the paper are to: (1) illustrate how the approach can be used with available data on standardized test scores, (2) assess the statistical precision of the approach and examine factors that affect this precision, and (3) briefly outline related issues to be addressed in future research. The paper is part of an ongoing effort to study the methodological properties of interrupted time-series analysis applied to data for a small number of periods (short time-series).

Introduction

For the past several decades, there has been considerable pressure to improve primary and secondary education in the U.S. This pressure is largely a response to sweeping economic changes caused by advanced technology and increased international competition. These changes have made it imperative for American workers to upgrade their skills, especially less-educated workers, whose standard of living has been declining steadily.[i1]

Responses to this need have increasingly come in the form of schoolwide restructuring programs.[ii2] Foremost among these are Henry Levin's Accelerated Schools Project,[iii3] Ted Sizer's Coalition of Essential Schools,[iv4] James Comer's School Development Program,[v5] and Robert Slavin's "Success for All" initiative.[vi6] Because such programs are designed to affect all students in a school, it is not possible to implement them for some students but not for others. Thus one cannot randomly assign individual students from the same school to a program group or control group.[vii7] This makes it difficult to determine what students would have achieved without restructuring (their "counterfactual") and thus how restructuring affected their achievement (its "impact").

Under certain conditions, however, it is possible to estimate the impacts of such programs by measuring the extent to which student achievement increased relative to its pre-program trend. This quasi-experimental design — referred to as interrupted time-series analysis — has been used in many different fields.[viii8] However, it has played little role in education research. This is probably due to concerns about: (1) the availability of consistent data over time, (2) the cost of obtaining data for individual students, (3) the limited precision of aggregate data, if individual data are not available, (4) confounding local events that can make it difficult to isolate the effect of restructuring, and (5) confounding changes in the mix of students which can make it difficult to interpret the effect of restructuring.[ix9]

It is possible, however, to overcome these problems under certain conditions. Thus, we are currently using interrupted time-series analysis as part of a major evaluation of the Accelerated Schools Project.[x10] From our experience, I believe that this approach can be used in other settings to help improve our understanding of how to promote educational excellence and equity.

Approach

Consider a school that launched a restructuring program five years ago. Assume that it has administered the same type of third-grade reading test for the past ten years and has maintained individual student records. Thus individual scores are available for a five-year baseline period (before the program was launched) and a five-year follow-up period (the launch year plus four subsequent years).[xi11]

One simple way to analyze these data is to plot the mean score for each year, as in Figure 1. A line fit through the means for years minus 5 through minus 1 represents the baseline trend. An extension of this line through the follow-up period (years zero through four) provides a forecast or prediction of what future mean scores would have been without restructuring (the counterfactual). The difference between the actual and predicted mean score for each follow-up year is thus, an estimate of the impact of restructuring for that year. I refer to these differences (labeled D_0 through D_4 in the figure) as "deviations from trend." These deviations illustrate the pattern of program impacts over time. For example, if the positive impact of a program is delayed several years because of the time required for its implementation, one would expect to see negligible deviations from trend at first, followed by positive ones in later years.

The preceding type of analysis has several important strengths. First, by adjusting for the baseline trend in test scores, the analysis controls for systemic changes in student performance over time, and thereby reduces the problem of "maturation". Second, by comparing future test scores to past test scores for several years — not just one — the analysis reduces the probability that baseline conditions represent an aberration, and thereby reduces the potential for "regression artifacts." Third, because the entire analysis can be summarized in a simple graph that can be understood readily by many different audiences, it provides an effective way to present impact findings. Fourth, because the analysis can be used with aggregate data on student test scores (discussed later) when individual test scores are not available, it has a broad range of potential applications. Fifth, because the analysis does not require detailed demographic data (although such data can be helpful if available), it can be simple to implement in practice. Sixth, because the analysis can be expressed as the following regression model, it provides a simple way to estimate program impacts and their standard errors.[xii12]

$$Y_i = A + B\ t_i + D_0\ F_{0i} + D_1\ F_{1i} + D_2\ F_{2i} + D_3\ F_{3i} + D_4\ F_{4i} + e_i \qquad \textbf{\textit{(1)}}$$

where:

Y_i = the test score for student i,
t_i = the test year for student i (ranging from - 5 through + 4 in the example),
F_{0i} = 1 if student i took the test in follow-up year zero and 0 otherwise,
F_{1i} = 1 if student i took the test in follow-up year one and 0 otherwise,
F_{2i} = 1 if student i took the test in follow-up year two and 0 otherwise,
F_{3i} = 1 if student i took the test in follow-up year three and 0 otherwise,
F_{4i} = 1 if student i took the test in follow-up year four and 0 otherwise,
e_i = the random *individual difference* in the score for student i, which is assumed to be independently and identically distributed, with a mean of zero and a variance of σ^2,
A and B = the intercept and slope of the pre-program trend respectively, and
D_0, D_1, $D_{2,}$ $D_{3,}$ and D_4 = deviations from trend (impact estimates) for follow-up years 0, 1, 2, 3, and 4 respectively.

The main potential weaknesses of the preceding analysis are "history" and "selection bias." A problem of history could arise if a major event — like a change in principal — coincided with the program and thereby provided an alternative explanation for the observed deviations from trend. A problem of selection bias could arise if a major shift in the mix of students coincided with the program and thereby made it difficult to interpret the observed deviations from trend.[xiii13]

Fortunately, by adding schools to the sample, it is possible to address these problems in two different ways. One way is to replicate the program in a number of different schools, conduct an interrupted time-series analysis in each, and pool the findings. This will reduce the "expected" overall influence of idiosyncratic local events (history) and post-program changes in student mix (selection).[xiv14] For this purpose it is best to have program schools that operate independently of each other (and thus are from different school districts) and that implement the program at different times (and thus are subject to different macro-historical conditions).[xv15]

A second approach — often referred to as a comparison series design — involves adding an interrupted time-series analysis for a comparison school which did not implement the program. By defining the baseline and follow-up periods for the comparison school in accord with those for the program school, and by computing the comparison school's deviations from *its own* baseline trend, a new estimate of the counterfactual is provided. This estimate predicts for any given follow-up year what the deviation from trend would have been without the program. The corresponding impact estimate is thus the difference between the program school's deviation from its trend and the comparison school's deviation from its trend.

However, because of the potential for idiosyncratic local events to occur, a study based only on one program school has little methodological protection, even if a comparison school is used. Thus, multiple program *and* comparison schools are highly desirable. Unfortunately, obtaining data for comparison schools (which have nothing to gain from cooperating and much to lose from invidious comparisons) can be much more difficult than obtaining data for program schools (that want to demonstrate success). This can become especially problematic for programs that are being tested in multiple school districts[xvi16] and thus require cooperation from many organizations to obtain comparison school data.

Thus, in order to use a comparison series design, it probably is necessary to limit the number of school districts involved — perhaps only to one. A program could then be tested in multiple schools per district, with a comparison school for each program school. In this context, it might be possible either to: (1) gain access to individual data for comparison schools with the help of school district officials; or (2) use aggregate test scores maintained by the school district for each school (discussed later).

Thus, in practice, there probably are two viable options for using interrupted time-series analysis to measure the impacts of education programs: (1) testing a program in multiple school districts *without* comparison schools or (2) testing a program in a small number of districts *with* comparison schools.

Both options can be used to estimate program impacts on student achievement in different subjects (reading, math, science, etc.), for students in different grades (e.g., third graders separately from sixth graders), and for outcomes other than test scores (e.g., attendance and retention in grade). Furthermore, both options can (and should) be combined with a detailed qualitative analysis of how a program was implemented, what influenced its success, and what else happened that might have affected student achievement.[xvii17]

Precision of the Estimates

One of the first questions to ask when considering the preceding approach is: "how much data are required?" More specifically: "how many baseline years are needed?" "how many schools are needed"? and "how large should the schools be?" Another important question to ask is: "how many follow-up years can be included in the analysis?" All of these questions are about factors that influence the statistical precision of program impact estimates, and thus to address them requires addressing the issue of precision.

A simple way to represent the precision of a research design is its "minimum detectable effect".[xviii18] Intuitively, this is the smallest impact that has a good chance of being identified if it actually exists. The smaller the minimum detectable effect is, the more precise the design is.

The first step in assessing the minimum detectable effect of a research design is to decide how impacts will be reported. A popular way to do so, especially for education research, is a measure called "effect size".[xix19] This is simply the impact in its original units (e.g., a scaled test score) divided by the standard deviation of the original measure for the population or sample of interest. Hence, effect sizes are measured in units of standard deviations. Thus, an effect size of 0.25 means a positive impact that is comparable in magnitude to 0.25 standard deviations. An effect size of - 0.40 means a negative impact that is comparable in magnitude to 0.40 standard deviations.

Although judgments about whether a specific effect size is large or small are ultimately arbitrary, some guidelines do exist. Many researchers use the rule of thumb proposed by Cohen (1977, 1988) which suggests 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 approach 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 0.00 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.

In the discussion which follows, I report the statistical precision of alternative interrupted time-series designs in terms of their minimum detectable effect size (MDES) and compare these findings to the preceding guidelines. To simplify the discussion, I focus on designs where all schools have the same number of baseline years and the same number of students per grade ("school size"). These examples provide an intuitive feel for how different factors affect statistical precision. In addition they provide a rough guide for making research design decisions. I begin with the simplest possible case and then extend the findings to more complex cases.

Program Schools Only, Without Cohort Differences

Equation 2 represents the minimum detectable effect size for a specific follow-up year, t_f, given an interrupted time-series design with program schools only (see Appendix A). By programming the equation in a spreadsheet, one can readily determine the precision of this design given different values for its parameters.

$$MDES(\bar{D}_f) = \frac{2.5}{\sqrt{mn}} \sqrt{1 + \frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}} \qquad \textit{\textbf{(2)}}$$

where:

$\bar{D}_f$ = the mean deviation from trend across all program schools in follow-up year, t_f,

m = the number of program schools,

n = the number of students per grade at each school,

T = the number of baseline years,

t_f = the follow-up year of interest,

$\bar{t}$ = the mean baseline year,[xx20]

$\sum_k (t_k - \bar{t})^2$ = the sum of squared variation of the *baseline* years around the mean baseline year.[xxi21]

Equation 2 is based on the interrupted time-series model in Equation 1, which assumes that the only source of random test score variance is *individual differences,* e_i. It does not account for potential random annual *cohort differences* which can increase the minimum detectable effect size (discussed later).

All of the relationships in Equation 2 are consistent with intuition. For example, increasing the number of schools, m, or choosing larger schools, n, reduces the minimum detectable effect size (which is inversely proportional to the square root of both parameters). Also, collecting data for more baseline years, T, reduces the minimum detectable effect size (both because of the role that T plays in the equation and because $\sum_k (t_k - \bar{t})^2$ increases with T). In all three cases, adding more data reduces the minimum detectable effect size and thereby increases statistical precision.

The last term in the equation, $(t_f - \bar{t})^2$, indicates that the minimum detectable effect size increases for later follow-up years (as the difference between t_f and $\bar{t}$ increases). In other words, impact estimates for later follow-up years are less precise than those for earlier years. This makes intuitive sense, because one should have less confidence in forecasts of the counterfactual for later years than for earlier years.[xxii22]

Table 1 summarizes the results of Equation 2 for a range of different baseline periods, follow-up years, and number of program schools, assuming 75 students per grade in each school. This is equivalent to three classes with 25 students each, which is typical of many urban elementary schools.

First consider the findings for one school. As can be seen, the minimum detectable effect size ranges from 0.39 to 0.90, and is 0.50 or larger in most cases. Hence, the precision of a one-school design is limited. This should not be surprising because the number of students each year is quite small (75 in the example). Unfortunately, however, it implies that studies which attempt to identify or compare impacts for individual schools are doomed to failure unless the programs they are evaluating produce consistently large impacts.

A second important finding in Table 1 is that statistical precision declines rapidly in later follow-up years. For a four-year baseline trend, the minimum detectable effect size almost doubles between follow-up year zero and follow-up year four; for a six-year trend it increases by roughly 50 percent. Hence, there is a limit to how long one can wait to identify program impacts.

A third important finding is that statistical precision increases substantially as one moves from a four-year baseline to a five-year baseline, especially for later follow-up years. This improvement is less pronounced as one moves from a five-year baseline to a six-year baseline. Thus, a five-year baseline might be a good compromise.

Now consider how the number of program schools affects precision. With 10 program schools the minimum detectable effect size approaches the range considered by Cohen (1977, 1988) and Lipsey (1990) to represent small impacts. For example, with five baseline years, the minimum detectable effect size ranges from 0.13 to 0.23. With 40 program schools and five baseline years, the minimum detectable effect size ranges from 0.07 to 0.11.

To examine more closely how the number of schools affects precision and to also examine the effect of "school size," consider the findings in Table 2. Note that these findings are for the second year after program launch, given a five-year baseline and no cohort differences.

The pattern of findings in the table reflect the simple fact that the minimum detectable effect size is inversely proportional to the square root of both the number of schools and school size. Hence, they exhibit "decreasing returns to scale" in terms of improving precision. For example, the marginal gain in precision from the first 20 schools is many times that for the next 20.

Differences in precision for school sizes of 50, 75, and 100 students per grade also exhibit decreasing returns to scale, but their pattern is less dramatic because the proportional change in school size represented in the table (from 50 to 100 students) is far less than that for the number of schools (from 1 to 40 schools).

Program Schools Only, With Cohort Differences

Now consider how the preceding findings change with the introduction of random cohort differences in student achievement caused by factors that can affect a whole grade at once, such as differences in class dynamics, changes in teaching staff, or revisions to a the test being used.[xxiii] [23]To account for these differences, the impact model requires an additional term, u_t , which is constant for all students from a school in a given year, but varies randomly over time. The model for a single school thus becomes:[xxiv][24]

$$Y_i = A + B\, t_i + D_0\, F_{0i} + D_1\, F_{1i} + D_2\, F_{2i} + D_3\, F_{3i} + D_4\, F_{4I} + u_t + e_i \qquad \textbf{\textit{(3)}}$$

where:

Y_i = the test score for student i,
t_i = the test year for student i (ranging from - 5 through + 4 in the example),
F_{0i} = 1 if student i took the test in follow-up year zero and 0 otherwise,
F_{1i} = 1 if student i took the test in follow-up year one and 0 otherwise,
F_{2i} = 1 if student i took the test in follow-up year two and 0 otherwise,
F_{3i} = 1 if student i took the test in follow-up year three and 0 otherwise,
F_{4i} = 1 if student i took the test in follow-up year four and 0 otherwise,
e_i = the random *individual difference* in the score for student i which is independently and identically distributed across students in a year with a mean of zero and a variance of σ^2,
u_t = the random *annual cohort difference* in the mean score for year t which is independently and identically distributed across years with a mean of zero and a variance of τ^2
A and B = the intercept and slope of the pre-program trend respectively, and

D_0, D_1, D_2, D_3, and D_4 = deviations from trend (impact estimates)
for follow-up years 0, 1, 2, 3 and 4 respectively.

The additional source of year-to-year random error reduces the stability of the baseline trend, which in turn increases its forecast error, and consequently reduces the precision of program impact estimates. Graphically this means that in Figure 1 the mean test score for each baseline year will vary more widely around the baseline trend. This, in turn, implies that the minimum detectable effect size will increase.

Equation 4 represents the minimum detectable effect size in the presence of annual cohort differences (for a derivation see Appendix B).

$$MDES(\bar{D}_f) = \frac{2.5}{\sqrt{m}} \sqrt{1/n + r/(1-r)} \sqrt{1 + \frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}} \qquad \textbf{\textit{(4)}}$$

where:

m = the number of program schools,
n = the number of students per grade at each school,
T = the number of baseline years,
t_f = the follow-up year of interest,

$\bar{t}$ = the mean baseline year,

$\sum_k (t_k - \bar{t})^2$ = the sum of squared variation of the baseline years around
the mean baseline year,

ρ = the intra-class correlation for between-year and within-year
random test score differences

First note the similarities between Equation 4 and Equation 2. In both equations, the minimum detectable effect size is inversely proportional to the square root of the number of schools, m. Also in both equations, the proportional effect of the number of baseline years (reflected by T and $\sum_k (t_k - \bar{t})^2$) and the proportional effect of the time elapsed between the follow-up year and the baseline period (t_f - $\bar{t}$) are the same.

However, the minimum detectable effect size in Equation 4 declines *by less* with respect to school size, n, than it does in Equation 2, because of the new term, (ρ/(1-ρ)). ρ is the annual intra-class correlation defined as the proportion of total random variation in test scores due to random cohort differences, or $\tau^2/(\tau^2+\sigma^2)$. Thus, ρ can take on values between zero and one. If ρ equals zero, there are no cohort differences — only individual differences. Hence, Equation 4 simplifies to Equation 2. If ρ equals one (which is virtually impossible) there are no individual differences — only cohort differences.

Researchers in other fields have assessed the magnitudes of intra-class correlations for various outcome measures and different types of groups. Foremost among this work is a series of papers by David Murray and his colleagues, who estimate intra-class correlations for use in the planning of community trials which test public health interventions. Because community trials randomly assign whole geographic areas (schools, school districts, cities, etc.) to treatment and control groups, a knowledge of their intra-class correlations is required to determine the precision of impact estimates.[xxv25]

Murray and Short (1995) used survey responses by 18- to 20-year-olds from 15 school districts in Wisconsin and Minnesota to estimate intra-class correlations *across districts* for measures of reported alcohol consumption. The overwhelming majority of their estimates were less than 0.01.[xxvi26] Murray et al. (1994) used findings from surveys conducted during the 1980s by eleven different researchers from across the country to estimate intra-class correlations *across schools* for measures of reported cigarette use. They focused mainly on three measures, which had mean intra-class correlations of 0.006, 0.011, and 0.019.[xxvii][27] Hannan et al. (1994) used survey data for six cities in upper Minnesota, South Dakota, and North Dakota over several years to estimate intra-class correlations *across cities and years* for a wide range of health outcomes related to heart disease. Their estimates "generally were in the range of 0.002 - 0.012."[xxviii][28]

Unfortunately, it is difficult to generalize these findings to the present analysis because intra-class correlations depend on how groups are defined and the groups used by previous researchers are not the same as annual student cohorts. Thus we need direct information about the intra-class correlation *across annual cohorts* for random test score variation (residuals) about a linear trend for a single school.

To begin to explore this issue empirically, I estimated values for ρ based on individual standardized math scores and reading scores for third-graders and sixth-graders from 25 elementary schools in Rochester, New York, during the four-year period from 1989-90 through 1992-93.[xxix][29] These intra-class correlations were estimated using individual residuals from a four-year linear trend for each school (see Appendix C). Separate estimates were obtained by grade and subject. Hence, there were 25 estimates for third-grade reading (one for each school), 25 estimates for third-grade math, 25 estimates for sixth-grade reading, and 25 estimates for sixth-grade math.

Table 3 summarizes the distributions of these estimates. As can be seen, most estimates were quite small and thus were consistent with estimates from previous research in other fields.

For reading, the intra-class correlation for the median school was near zero for third-graders and below 0.01 for sixth-graders. Even schools at the 75th percentile had a relatively small intra-class correlation: 0.01 for third-graders and 0.02 for sixth graders. Thus, three out of four schools had intra-class correlations that were 0.01 or smaller for third-graders and 0.02 or smaller for sixth-graders.

For math, the intra-class correlations were somewhat larger, but in most cases they were still fairly small. The median school had a value 0.02 for both third-graders and sixth-graders. However, schools at the 75th percentile had values of 0.04 and 0.07, respectively. Hence, for some schools the intra-class correlation for math was substantial, but for most it was still within the range of values observed in other fields.

To illustrate how the intra-class correlation can affect the precision of impact estimates from an interrupted time-series analysis, Equation 4 was used to compute values of the minimum detectable effect size for intra-class correlations of 0.01, 0.03, and 0.05. Several tentative conclusions emerge from these findings, which are presented in Table 4.

First, the intra-class correlation makes a big difference in precision. For example, the minimum detectable effect size for 10 schools and 75 students per grade is 0.23, 0.32, and 0.39 for ρ equal to 0.01, 0.03, and 0.05 respectively. Thus as ρ increases a little, the minimum detectable effect size increases a lot.

Second, accounting for ρ makes it even more clear that meaningful impact estimates for individual schools probably are not feasible. The minimum detectable effect size for one school equals 0.74, 1.01, or 1.24 when ρ equals 0.01, 0.03, or 0.05, respectively. These imply very large impacts, which are quite unlikely to be achieved in practice.

Third, as ρ increases, the influence of school size on the precision of impact estimates declines appreciably. For example, with ρ equal to 0.01, the minimum detectable effect size for one school with 50 versus 100 students per grade is 0.83 versus 0.68; for ρ equal to 0.05, the corresponding minimum detectable effect sizes are 1.30 and 1.20, respectively.

Fourth, and perhaps most important, is that between 10 and 20 program schools might provide adequate statistical precision for an impact study. For values of ρ around 0.01, which is what the Rochester findings suggest, the minimum detectable effect size for 10 program schools ranges from 0.26 to 0.22 and those for 20 schools range from 0.19 to 0.15. Even for values of ρ around 0.03 (which is beyond most estimates of intra-class correlations obtained to date), the minimum detectable effect size for 10 schools ranges from 0.34 to 0.31 and that for 20 schools ranges from 0.24 to 0.22.

Adding Comparison Schools

Now consider what happens to the precision of impact estimates if we add comparison schools to the research design. Recall that the purpose of adding comparison schools is to help guard against problems of history and selection. Hence, they strengthen the research design in this regard. However, this additional methodological strength comes at the cost of reduced statistical precision. Fortunately this reduction is not insurmountable, so that using a comparison series design can be justified if appropriate data are available.

For a design with: one comparison school per program school, all schools having the same number of students per grade, and all schools having the same number of baseline years, one can simply multiple the minimum detectable effect size for any number of program schools by $\sqrt{2}$ (approximately 1.414) to obtain its counterpart for a comparison series design (see Appendix D). This holds both for Equations 2 and 4 and, thus, for all findings in Tables 1, 2, and 4.

For example, recall that the minimum detectable effect size in Table 4 for 10 schools, 75 students per grade per school, and ρ equal to 0.01 was 0.23. Adding a comparison school for each program school increases the minimum detectable effect size to approximately 1.414(0.23) or 0.33, which is still in range of small to moderate impacts.

Further Research

This paper has argued that interrupted time-series analysis has a potentially important role to play in the evaluation of programs that are intended to affect whole-school change. Indeed the findings presented above suggest that baseline test data from 10 to 20 program schools for five or six years might be adequate to provide defensible impact estimates for three to five follow-up years.

Therefore I believe that additional research is warranted to further explore the methodological properties of this approach and its feasibility in practice. Some of this research is underway currently as part of our evaluation of the national Accelerated Schools Project. In this final section, I briefly outline what has been done and what else is planned. Future papers will report on this research.

Using Aggregate Data

As discussed earlier, one major concern about using interrupted time-series analysis for research in education is the potential lack of adequate data. However, this problem could be reduced substantially if it were possible to use average annual test scores. School districts often publish such data by school, and data are becoming increasingly available through the Internet.

Fortunately, it may be possible to use this aggregate data because annual average test scores, like those plotted in Figure 1, can provide impact estimates that are identical to those obtained by estimating Equation 1 or Equation 3 from data on individual test scores.[xxx30] In addition aggregate data can provide valid estimates of the standard error of impact estimates. What is lost, however, when moving from individual data to aggregate data, is a certain amount of statistical precision due to the limited number of aggregate observations, and thus the limited number of degrees of freedom available to estimate standard errors. We have developed a method for comparing the minimum detectable effect size of aggregate versus individual data, and our preliminary findings suggests that five or six years of aggregate baseline data plus subsequent aggregate follow-up data may be adequate for an interrupted time-series analysis with 10 to 20 program schools.

Pooling Across Schools

As noted earlier, one school is not enough to estimate the impacts of an education program — for any methodology, not just the present one. Thus pooling findings across a number of schools is essential. We therefore are exploring issues that arise when doing so as part of the Accelerated Schools evaluation.

In particular, we are considering how best to pool findings across different schools which used different tests, although each school used the same test over time. To do so, we will convert all scores for each school into z scores defined in terms of its baseline mean and standard deviation. This will convert all impact estimates to an effect size metric, which in turn will facilitate two ways of pooling: (1) computing separate impact estimates for each school and pooling the estimates or (2) pooling the data and computing one estimate.

Assessing Equity Impacts

This paper has focused on *mean* student achievement, implicitly as a measure of educational excellence. But educational excellence can be attained in different ways which have different equity implications. For example, programs that emphasize enrichment for “gifted and talented” students can increase mean achievement by increasing the test scores of students with the strongest backgrounds. This, in turn, can broaden the gap between students with the strongest backgrounds and those with the weakest. In contrast, initiatives which pay special attention to “at risk” students — like Accelerated Schools — can raise mean achievement while reducing the gap between students at the top and students at the bottom of the achievement distribution.

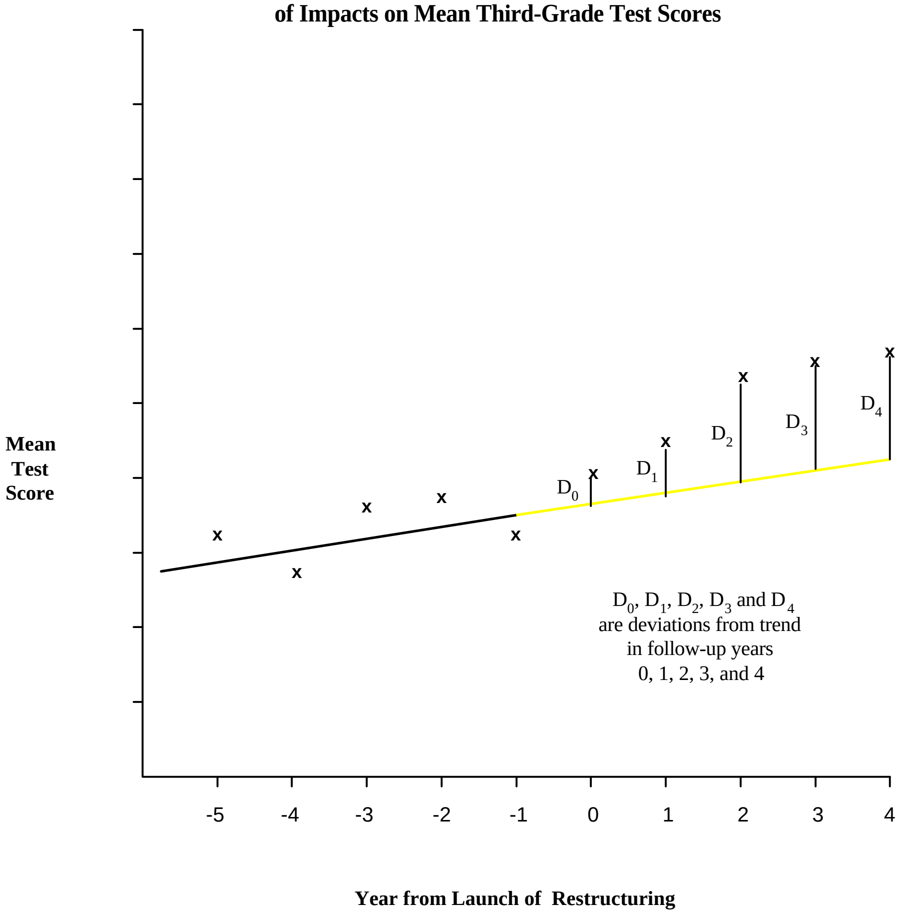
Therefore one way to assess the equity implications of an educational program is to measure its effect on the *standard deviation* of student test scores. Using interrupted time-series analysis we plan to do this for the Accelerated Schools evaluation. For every school in the sample, we will estimate how the standard deviation for each follow-up year differed (deviated) from its pre-program trend.

Exploring Alternative Trends

The present paper focused only on linear baseline trends. This was done both because linear functions are used widely for many types of research, and because a preliminary analysis of time-series data for Rochester test scores suggests that linear trends may be appropriate. However, we also plan to examine several alternative functional forms. In particular, we will consider forecasting future mean test scores from: (1) the overall mean baseline score (which assumes no systemic change over time), (2) a quadratic function of time (which allows the change in test scores to accelerate or decelerate over time), and (3) a logarithmic function of time (which allows the change in test scores to decelerate over time).

Figure 1

A Hypothetical Interrupted Time-Series Analysis of Impacts on Mean Third-Grade Test Scores



Note: Years -5 through -1 in the figure represent the pre-launch baseline period, and the line through the mean test scores for these years reprents the pre-launch trend. Year 0 represents the launch year and years 1 through 4 represent the post-launch years.

Table 1

Minimum Detectable Effect Size
By Baseline Period, Follow-up Year and Number of Program Schools
(for 75 students per grade, no comparison schools and no cohort differences)

Follow-up	**Baseline Period**		
Year	**Four Years**	**Five Years**	**Six Years**
	1 Program School		
Zero	0.46	0.42	0.39
One	0.56	0.48	0.44
Two	0.67	0.56	0.49
Three	0.78	0.63	0.55
Four	0.90	0.71	0.60
	10 Program Schools		
Zero	0.14	0.13	0.13
One	0.18	0.15	0.14
Two	0.21	0.18	0.16
Three	0.25	0.20	0.17
Four	0.28	0.23	0.19
	40 Program Schools		
Zero	0.07	0.07	0.06
One	0.09	0.08	0.07
Two	0.11	0.09	0.08
Three	0.12	0.10	0.09
Four	0.14	0.11	0.10

Table 2

Minimum Detectable Effect Size
By Number of Program Schools and Students Per Grade
(for follow-up year 2, with 5 baseline years and no cohort differences)

Number of	**Students per Grade**		
Program Schools	**50 Students**	**75 Students**	**100 Students**
1 School	0.68	0.56	0.48
5 Schools	0.30	0.25	0.22
10 Schools	0.22	0.18	0.15
20 Schools	0.15	0.12	0.11
30 Schools	0.12	0.10	0.09
40 Schools	0.11	0.09	0.08

Table 3

Distribution of Estimated Intra-Class Correlations
For Individual Residuals from a Four-Year Linear Time-Trend
For 25 Elementary Schools from Rochester, New York[a]

	Third Grade		**Sixth Grade**	
	Reading	**Math**	**Reading**	**Math**
25th percentile	0.00	0.00	0.00	0.00
50th percentile (median)	0.00	0.02	<0.01	0.02
75th percentile	0.01	0.04	0.02	0.07

[a] For standardized reading and math tests administered to third-graders and sixth-graders each year between 1989-90 and 1992-93.

Table 4

Minimum Detectable Effect Size
By Number of Program Schools,
Students Per Grade, and Intra-Class Correlation
(for follow-up year 2, with 5 baseline years and cohort differences)

Number of Program Schools	**Students per Grade**		
	50 Students	**75 Students**	**100 Students**
	Intra-Class Correlation = 0.01		
1 School	0.83	0.74	0.68
5 Schools	0.37	0.33	0.31
10 Schools	0.26	0.23	0.22
20 Schools	0.19	0.17	0.15
30 Schools	0.15	0.13	0.12
40 Schools	0.13	0.12	0.11
	Intra-Class Correlation = 0.03		
1 School	1.09	1.01	0.97
5 Schools	0.49	0.45	0.44
10 Schools	0.34	0.32	0.31
20 Schools	0.24	0.23	0.22
30 Schools	0.20	0.19	0.18
40 Schools	0.17	0.16	0.15
	Intra-Class Correlation = 0.05		
1 School	1.30	1.24	1.20
5 Schools	0.58	0.55	0.54
10 Schools	0.41	0.39	0.38
20 Schools	0.29	0.28	0.27
30 Schools	0.24	0.23	0.22
40 Schools	0.21	0.20	0.19

Appendix A

The Minimum Detectable Effect Size for Program Schools Only, Without Cohort Differences

This appendix derives the minimum detectable effect size (MDES) for a single follow-up year from an interrupted time-series design with program schools only and no cohort differences (Equation 2 in the paper). Some of the notation and conventions in this appendix differ from those in the paper, because the paper was simplified to facilitate the discussion.

Consider an interrupted time-series analysis for a single program school with a T-year baseline period that begins in year t_1 and ends in year $t_{T.}$ Thus a five-year baseline period would run from t_1 to t_5, and the follow-up period would comprise a program launch year in t_6, a first follow-up year in t_7, and so on. Assume that n students take the test each year.

The program impact for a single follow-up year, t_f, can be estimated as D_f from the following regression model using data for the baseline period and the follow-up year.[xxxi]

$$Y_{ki} = \quad A + B\,t_{ki} + D_f\,F_{fki} + e_{ki} \qquad \textbf{\textit{(A1)}}$$

where:

Y_{ki} = the test score for student i in year k (*recall that different students take the test each year*),[xxxii32]
t_{ki} = the year indicator for student i in year k (equal to k),
F_{fki} = one for students who took the test in follow-up year t_f and zero otherwise,
A = the intercept of the baseline trend,
B = the slope of the baseline trend,
D_f = the deviation from trend (program impact) in follow-up year t_f, and
e_{ki} = the random individual difference for student i in year k, which is independently and identically distributed with a mean of zero and a constant variance of σ^2.

Deriving the Variance of the Impact Estimate

The estimated deviation from trend $\hat{D}_f$ is equal to the mean test score for the follow-up year $\bar{Y}_f$ minus the predicted test score for that year $\hat{Y}_f$. In other words:

$$\hat{D}_f = \bar{Y}_f - \hat{Y}_f \qquad \textbf{\textit{(A2)}}$$

where:

$$\hat{Y}_f = \hat{A} + \hat{B}t_f \qquad \textbf{\textit{(A3)}}$$

Thus

$$VAR(\hat{D}_f) = VAR(\bar{Y}_f) + VAR(\hat{Y}_f) \qquad \textbf{\textit{(A4)}}$$

Given the properties of e_{ki} in Equation 1,

$$VAR(\bar{Y}_f) = \frac{s^2}{n} \qquad \textbf{\textit{(A5)}}$$

To obtain $VAR(\hat{Y}_f)$ first note that:

$$VAR(\hat{Y}_f) = VAR(\hat{A} + \hat{B}t_f) \qquad \textbf{\textit{(A6)}}$$

and thus

$$VAR(\hat{Y}_f) = VAR(\hat{A}) + t_f^{\,2} VAR(\hat{B}) + 2t_f COV(\hat{A}, \hat{B}) \qquad \textbf{\textit{(A7)}}$$

Applying expressions for the variances and covariance of the intercept and slope of a bivariate regression (Pindyck and Rubinfeld, 1998, pp. 63-64) and simplifying, yields:

$$VAR(\hat{A}) = \frac{s^2 \sum_k \sum_i t_{ki}^{\,2}}{nT \sum_k \sum_i (t_{ki} - \bar{t})^2} = \frac{s^2 n \sum_k t_k^{\,2}}{nTn \sum_k (t_k - \bar{t})^2} = \frac{s^2}{n}\left[\frac{\sum_k t_k^{\,2}}{T \sum_k (t_k - \bar{t})^2}\right] \qquad \textbf{\textit{(A8)}}$$

$$t_f^{\,2} VAR(\hat{B}) = \frac{t_f^{\,2} s^2}{\sum_k \sum_i (t_{ki} - \bar{t})^2} = \frac{s^2}{n}\left[\frac{t_f^{\,2}}{\sum_k (t_k - \bar{t})^2}\right] \qquad \textbf{\textit{(A9)}}$$

$$-2t_f COV(\hat{A}, \hat{B}) = \frac{-2t_f \bar{t} s^2}{\sum_k \sum_i (t_{ki} - \bar{t})^2} = \frac{s^2}{n}\left[\frac{-2t_f \bar{t}}{\sum_k (t_k - \bar{t})^2}\right] \qquad \textbf{\textit{(A10)}}$$

where $\sum_k = \sum_{k=1}^{T}$ and $\sum_i = \sum_{i=1}^{n}$

Substituting Equations A8-A10 into Equation A7 and simplifying yields:

$$VAR(\hat{Y}_f) = \frac{s^2}{n}\left[\frac{\sum_k t_k^{\,2}}{T \sum_k (t_k - \bar{t})^2}\right] + \frac{s^2}{n}\left[\frac{t_f^{\,2}}{\sum_k (t_k - \bar{t})^2}\right] - \frac{s^2}{n}\left[\frac{2t_f \bar{t}}{\sum_k (t_k - \bar{t})^2}\right]$$

$$= \frac{s^2}{n}\left[\frac{\sum_k t_k^{\,2}}{T \sum_k (t_k - \bar{t})^2} + \frac{t_f^{\,2}}{\sum_k (t_k - \bar{t})^2} - \frac{2t_f \bar{t}}{\sum_k (t_k - \bar{t})^2}\right] \qquad \textbf{\textit{(A11)}}$$

Expanding $\sum_k t_k^2$ and simplifying yields:

$$VAR(\hat{Y}_f) = \frac{s^2}{n}\left[\frac{\sum_k (t_k - \bar{t})^2 + T\bar{t}^2}{T\sum_k (t_k - \bar{t})^2} + \frac{t_f^2}{\sum_k (t_k - \bar{t})^2} - \frac{2t_f\bar{t}}{\sum_k (t_k - \bar{t})^2}\right]$$

$$= \frac{s^2}{n}\left[\frac{1}{T} + \frac{\bar{t}^2 + t_f^2 - 2t_f\bar{t}}{\sum_k (t_k - \bar{t})^2}\right]$$

$$= \frac{s^2}{n}\left[\frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}\right] \qquad \textbf{\textit{(A12)}}$$

Substituting Equations A12 and A5 into Equation A4 and simplifying yields:

$$VAR(\hat{D}_f) = VAR(\bar{Y}_f) + VAR(\hat{Y}_f)$$

$$= \frac{s^2}{n} + \frac{s^2}{n}\left[\frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}\right]$$

$$= \frac{s^2}{n}\left[1 + \frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}\right] \qquad \textbf{\textit{(A13)}}$$

Equation A13 represents the standard error of forecast for a bivariate regression (Pindyck and Rubinfeld, 1998, p. 208) applied to mean annual test scores for the baseline period and the follow-up year.

Generalizing to Multiple Program Schools

Extending Equation A13 to a mean impact estimate, $\bar{D}_f$, for m program schools with the same grade size, number of baseline years and individual test variance, σ^2, yields:

$$VAR(\bar{D}_f) = \frac{s^2}{mn}\left[1 + \frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}\right] \qquad \textbf{\textit{(A14)}}$$

Obtaining the Minimum Detectable Effect

Bloom (1995) demonstrates that the minimum detectable effect of an impact estimator is a simple multiple of its standard error. For a one-tail hypothesis test at the 0.05 level with 80 percent power, the multiple is 2.5. Thus:

$$MDE(\bar{D}_f) = 2.5\sqrt{VAR(\bar{D}_f)} = \frac{2.5s}{\sqrt{mn}}\sqrt{1+\frac{1}{T}+\frac{(t_f-\bar{t})^2}{\sum_k (t_k-\bar{t})^2}} \qquad \textbf{\textit{(A15)}}$$

Obtaining the Minimum Detectable Effect Size

Dividing Equation A15 by the standard deviation of individual test scores each year, σ, yields the minimum detectable effect size. Therefore:

$$MDES(\bar{D}_f) = \frac{2.5}{\sqrt{mn}}\sqrt{1+\frac{1}{T}+\frac{(t_f-\bar{t})^2}{\sum_k (t_k-\bar{t})^2}} \qquad \textbf{\textit{(A16)}}$$

Appendix B

The Minimum Detectable Effect Size for Program Schools Only With Cohort Differences

This appendix derives the minimum detectable effect size for an interrupted time-series model with random year-to-year cohort differences. To do so, first note that Equation A14 from Appendix A (without cohort differences) can be re-expressed as follows:

$$VAR(\bar{D}_f) = \frac{s^2}{mn}\left[1+\frac{1}{T}+\frac{(t_f-\bar{t})^2}{\sum_k (t_k-\bar{t})^2}\right]$$

$$= \frac{VAR(\bar{e}_k)}{m}\left[1+\frac{1}{T}+\frac{(t_f-\bar{t})^2}{\sum_k (t_k-\bar{t})^2}\right] \qquad \textit{(B1)}$$

where $\bar{e}_k$ is the mean value of e_{ki} for year k and $VAR(\bar{e}_k)$ is the year-to-year variance of this mean.

Now consider the interrupted time-series model for one follow-up year, t_f, and a random year-to-year cohort difference, u_k:

$$Y_{ki} = A + B\, t_{ki} + D_f\, F_{fki} + u_k + e_{ki} \qquad \textit{(B2)}$$

where

Y_{ki} = the test score for student i in year k,
t_{ki} = the year indicator for student i in year k (equal to k),
F_{fki} = one for students who took the test in follow-up year t_f and zero otherwise,
A = the intercept of the baseline trend,
B =the slope of the baseline trend,
D_f = the deviation from trend (program impact) in follow-up year t_f, and
e_{ki} = the random individual difference for student i in year k, which is independently and identically distributed across students, with a mean of zero and a variance of σ^2,
u_k = the random annual cohort difference for year k, which is constant for all students in year k and is independently and identically distributed across years, with a mean of zero and a variance of τ^2.

The year-to-year variance of mean annual test scores around the trend for this model is:

$$VAR(u_k+\bar{e}_k) = t^2 + \frac{s^2}{n} \qquad \textit{(B3)}$$

Re-expressing Equation B3 in terms of σ^2 and the intra-class correlation, ρ, defined as $\tau^2/(\tau^2+\sigma^2)$, yields:

$$VAR(u_k + \bar{e}_k) = s^2\left(\frac{1}{n} + \frac{r}{1-r}\right) \tag{B4}$$

Replacing VAR($\bar{e}_k$) in Equation B1 with the expression for $VAR(u_k + \bar{e}_{ki})$ in Equation B4 and simplifying yields:

$$VAR(\bar{D}_f) = \frac{VAR(u_k + \bar{e}_k)}{m}\left[1 + \frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}\right]$$

$$= \frac{s^2(1/n + r/(1-r))}{m}\left[1 + \frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}\right] \tag{B5}$$

Converting VAR($\bar{D}_f$) to it counterpart expression for the minimum detectable effect size, as was done in Appendix A, thus yields:[xxxiii33]

$$MDES(\bar{D}_f) = \frac{2.5}{\sqrt{m}}\sqrt{1/n + r/(1-r)}\sqrt{1 + \frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}} \tag{B6}$$

Equation B6 is presented as Equation 4 in the paper.

Appendix C

Estimating the Intra-Class Correlation for Annual Cohort Effects

This appendix describes how estimates were obtained for the intra-class correlation due to annual cohort effects. Estimates were based on individual scores for third-grade and sixth-grade math and reading tests for 25 elementary schools in Rochester, New York, during the four years from 1989-90 through 1992-93. Student scores were from the Pupil Evaluation Program (PEP) test, which is administered annually throughout New York State. The average number of third graders per year per school in the sample ranged from 29 to 121, with a mean of 71. The corresponding number of sixth graders ranged from 21 to 96, with a mean of 54.

The model used for this analysis was:

$$Y_{ki} = A + B\, t_{ki} + u_k + e_{ki} \qquad \textit{(C1)}$$

where:

Y_{ki} = the test score for student i in year k,
t_{ki} = the year indicator for student i in year k (equal to k),
A = the intercept of the baseline trend,
B = the slope of the baseline trend,
u_k = the random annual cohort difference for year k, which is constant for all students in year k and is independently and identically distributed across years, with a mean of zero and a variance of τ^2,
e_{ki} = the random individual difference for student i in year k, which is independently and identically distributed across students, with a mean of zero and a variance of σ^2.

This model for a single school specifies a linear trend in test scores over time with an individual stochastic component, e_{ki}, and a cohort stochastic component, u_k. The intra-class correlation, ρ, for these two stochastic components is thus $\tau^2/(\tau^2+\sigma^2)$.

Estimates of the intra-class correlation were obtained by school, grade and subject. For each school, ordinary least squares (OLS) was used to estimate the regression in Equation C1 from the four years of available scores for a particular grade and subject. Next τ^2 and σ^2 were estimated from the residuals of this regression using SAS VARCOMP.[xxxiv34] These findings were used to compute an intra-class correlation for each school which, in turn, were ranked from lowest to highest. Summary statistics in Table 3 of the paper were obtained from the resulting distribution for each grade and subject.

Appendix D

The Minimum Detectable Effect Size for Program and Comparison Schools With or Without Cohort Differences

Perhaps the simplest way to think about adding comparison schools to the basic interrupted time-series design is to pool all deviations from trend for a specific follow-up year, t_f, for program schools (as we have been doing) and then to separately pool all deviations from trend for comparison schools for each corresponding follow-up year. For each follow-up year this would provide a mean deviation from trend for program schools, $\bar{D}_{fp}$, and a mean deviation from trend for comparison schools, $\bar{D}_{fc}$. The *difference* between these two deviations from trend is the program impact estimate. In the simplest case, with one comparison school for each program school, and each school having the same number of students per grade, n, and the same number of baseline years, T, the variance of the impact estimate is:[xxxv35]

$$VAR(\bar{D}_{fp} - \bar{D}_{fc}) = VAR(\bar{D}_{fp}) + VAR(\bar{D}_{fc}) \qquad \textbf{\textit{(C1)}}$$

For research design purposes, one can assume that the variance components for program schools are roughly the same as those for comparison schools. Thus

$$VAR(\bar{D}_{fp} - \bar{D}_{fc}) = 2VAR(\bar{D}_{fp}) \qquad \textbf{\textit{(C2)}}$$

This implies that the standard error of an impact estimate which includes comparison schools equals the $\sqrt{2}$ times the standard error of an impact estimate without comparison schools. Thus the minimum detectable effect size with comparison schools equals the $\sqrt{2}$ times the minimum detectable effect size without comparison schools. This finding holds both with or without annual cohort differences.

Notes

[1] See, for example, Levy (1998), Blank (1997), and Burtless (1990).

[2] Murphy (1993) provides a brief review of these programs.

[3] See, for example, Levin (1993).

[4] See, for example, Prestine (1993).

[5] See, for example, Barnett (1996).

[6] See, for example, Slavin et al. (1993).

[7] It might be possible to randomly assign students from a pool of volunteers to a "program school" that has restructured (see Cave and Kagehiro, 1995). However, this is only possible under special circumstances. In addition, for some initiatives — such as a new curriculum component — it might be possible to randomly assign whole schools to a program or a control group. Raudenbush (1997) provides a theoretical discussion of the statistical properties of this "cluster random assignment" approach; and Bloom, Bos, and Lee (1998) provide an empirical analysis of these properties. In addition, Cook et al. (forthcoming) used the approach to evaluate Comer's Student Development Program in Prince George's County Maryland, and Cook et al. (1999) used it to evaluate Comer's program in Chicago. However, given the extensive "buy-in" required to restructure a school, it will not always be possible to use cluster random assignment in this context.

[8] Shadish (in preparation) provides a comprehensive review of the interrupted time-series literature. Campbell and Stanley (1966) and Cook and Campbell (1979) are perhaps the most widely cited sources on the topic.

[9] My thanks to Bob Granger of MDRC for raising these issues.

[10] See Ham and Rock (1999).

[11] These data represent "moving cross-sections" for different third-grade cohorts each year. Hence, they represent different students in different years (with the exception of those who are held back and consequently retested).

[12] The parameters in Equation 1 are *identical* to their counterparts in Figure 1 when the number of students per year is constant. If the number of students varies over time, an appropriately weighted analysis of the annual means in Figure 1 will reproduce the parameter estimates for Equation 1.

[13] Selection bias exists if the program did not cause the shift in student mix. However, selection does not create a problem of *bias* if the shift in student mix was caused by the program. It does, however, create a problem of *interpretation*. Without more data and/or further assumptions one cannot distinguish between: (1) the change in test scores due to the change in student mix caused by the program, or (2) the change in test scores due to the change in individual achievement caused by the program. In both cases, however, the change in test scores was caused by the program.

[14] Idiosyncratic events and major changes in student mix should be distributed randomly across schools which operate independently of each other. Some of these changes will cause student achievement to be above the baseline trend, and others will cause student achievement to be below the baseline trend. Across a large number of schools, however, the average deviation from trend should be zero.

[15] We are basing our evaluation of the Accelerated Schools Project on the experiences of eight to ten schools from six to seven states that launched their programs between 1990-91 and 1993-94. These schools have consistent test data for a ten-year analysis period and were judged by staff from the national Accelerated Schools Project to have reached "mature" acceleration.

[16] For the Accelerated Schools evaluation, which involved eight to ten school districts in six to seven states, we found it very difficult to obtain comparison school data. We thus are using a multiple program school design without comparison schools.

[17] Ham and Rock (1999) describe the integrated quantitative and qualitative approach being used to evaluate Accelerated Schools.

[18] Bloom (1995) illustrates how the minimum detectable effect of an estimator is a simple multiple of its standard error. This multiple depends on three factors: (1) the desired level of statistical significance, (2) the desired level of statistical power, and (3) whether a one-tail or two-tail test is being used.

[19] Effect size measures are especially popular for meta-analyses which pool impact estimates across different outcomes and metrics. For example, see Hedges and Olkin (1985) or Rosenthal (1991).

[20] For example, the mean value of five baseline years (-5 through -1) is -3.

[21] For example, the sum of squared variation for five baseline years (-5 through -1) is 10.

[22] The finding that precision declines as one forecasts further beyond the data used to estimate a model is well known in econometrics (e.g., see Pindyck and Rubinfeld, 1998, pp. 204-209).

[23] Throughout this paper I assume that the same test is used over time for a given school, although different versions of the same test may be used. This constraint is imposed on the design to avoid shifts in scores that occur when a new test is implemented. Linn, 1998, pp. 11-12, illustrates how scores fall precipitously when a new test is implemented and then rebound over time as teachers learn how to prepare students better (i.e., teach to the test). To evaluate the Accelerated Schools Project, we therefore only selected schools that used the same test for at least five years before and after launching their program.

[24] Equation 3 is a "random-effects" model. Such models are used in: (1) the econometrics literature on panel data (see, for example, Greene, 1997, Chapter 14); (2) the statistics literature on analysis of variance (see, for example, Hays, 1973, Chapter 13); (3) the meta-analysis literature on pooling findings across studies (see, for example, Raudenbush, 1994), and (4) the literature on "hierarchical linear models" (see, for example, Bryk and Raudenbush, 1992, Chapter 2).

[25] See Raudenbush (1997) for a discussion of the statistical issues involved.

[26] See Murray and Short (1995) Tables 2-8, pp. 685-691.

[27] See Murray et al. (1994), p. 1042.

[28] See Hannan et al. (1994), abstract, p. 88.

[29] My thanks to Michelle Moser and Steve Caso for making these data available.

[30] This represents a special case of aggregation where the value of each independent variable (the year indicator, t_i, and the dummy variables for each follow-up year, F_{0i}, F_{1i}, F_{2i}, F_{3I}, and F_{4i}) is the same for all members of the same aggregate unit (a cohort of students for a given year). Kmenta (1971), pp. 322-325, demonstrates that when this condition is met, the point estimates from an aggregate regression are identical to those from the model for individuals.

[31] To simplify the discussion, Equation A1 includes all baseline years but only one follow-up year. Equation 2 in the paper generalizes the model to include all follow-up years.

[32] Thus, the i^{th} student in year one is different from the i^{th} student in year two, and so on.

[33] Equation B6 thus represents the minimum detectable effect size for a one-tail hypothesis test at the 0.05 significance level with 80 percent power.

[34] See SAS Institute Inc. (1989).

[35] This assumes that random variation in tests scores for program schools and comparison schools are independent of each other.

References

Barnett, W. Steven (1996) "Economics of School Reform: Three Promising Models" in Helen F. Ladd ed. *Holding Schools Accountable: Performance-Based Reform in Education* (Washington DC: The Brookings Institution) pp. 299-326.

Blank, Rebecca M. (1997) *It Takes a Nation: A New Agenda for Fighting Poverty* (New York: The Russell Sage Foundation).

Bloom, Howard S., Johannes M. Bos and Suk-Won Lee (1998) "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs" (New York: Manpower Demonstration Research Corporation, April).

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

Bryk, Anthony S. and Stephen W. Raudenbush (1992) *Hiearchical Linear Models: Applications and Data Analysis Methods* (Newbury Park, CA: Sage Publications).

Burtless, Gary (1990) *A Future of Lousy Jobs: The Changing Structure of U.S. Wages* (Washington, DC: The Brookings Institution).

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

Cave, George and Susie Kagehiro (1995) "Accelerated Middle Schools: Assessing the Feasibility of a Net Impact Evaluation" (New York: Manpower Demonstration Research Corporation, June).

Cohen, J. (1977) *Statistical Power Analysis for the Behavioral Sciences*, rev. ed. (New York: Academic Press).

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

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

Cook, Thomas, Farah-naaz Habib, Meredith Phillips, Richard A. Settersten, Shobha C. Shagle and Serdar M. Degirmencioglu (forthcoming) "Comer's School Development Program in Prince George's County, Maryland: A Theory-Based Evaluation", *American Educational Research Journal* (Summer, 1999).

Cook, Thomas, H. David Hunt and Robert F. Murphy (1999) "Comer's School Development Program in Chicago: A Theory-Based Evaluation" (Northwestern University, Institute for Policy Research Working Paper).

Greene, William H. (1997) *Econometric Analysis*, 3rd edition (Upper Saddle River, New Jersey: Prentice Hall).

Ham, Sandra and JoAnn L. Rock (1999) "The Accelerated Schools Evaluation: Integrating Quantitative and Qualitative Measures of Whole-School Reform" (New York: Manpower Demonstration Research Corporation, April).

Hannan, Peter J., David Murray, David R. Jacobs, Jr., and Paul McGovern (1994) "Parameters to Aid in the Design and Analysis of Community Trials: Intraclass Correlations from the Minnesota Heart Health Program" *Epidemiology*, Vol. 5, No. 1, January, pp. 88-94.

Hays, William L. (1973) *Statistics for the Social Sciences,* 2nd edition (New York: Holt, Rinehart and Winston, Inc.)

Hedges, Larry V. and Ingram Olkin (1985) *Statistical Methods for Meta-Analysis* (New York: Academic Press).

Kmenta, Jan (1971) *Elements of Econometrics* (New York: MacMillan Publishers).

Levin. Henry M. (1993) "Learning from Accelerated Schools" in James H. Block, Susan T. Everson and Thomas R. Guskey eds. *Selecting and Integrating School Improvement Programs* (New York: Scholastic Books).

Levy, Frank (1998) *The New Dollars and Dreams: American Incomes and Economic Change* (The Russell Sage Foundation).

Linn, Robert L. (1998) “Assessments and Accountability” (University of Colorado at Boulder: Center for Research on Evaluation, Standards and Student Testing, April).

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

Moser, Sir Claus and G. Kalton (1972) “Clustering and Multi-Stage Sampling” in *Survey Methods in Social Investigation* (New York: Basic Books), pp. 100-111.

Murphy, Joseph (1993) “Restructuring: In Search of a Movement” in Joseph Murphy and Philip Hallinger eds. *Restructuring Schooling: Learning from Ongoing Efforts* (Newbury Park, CA: Corwin Press Inc.) pp. 1-31.

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*, Vol. 56, No. 6, November, pp. 681-693.

Murray, David M., Brenda L. Rooney, Peter J. Hannan, Arthur V. Peterson, Dennis V. Ary, Anthony Biglan, Gilbert J. Botvin, Richard I. Evans, Brian R. Flay, Robert Futterman, J. Greg Getz, Pat M. Marek, Mario Orlandi, Mary Ann Pentz, Cheryl L. Perry and Steven P. Schinke (1994) “Intra-class Correlation among Common Measures of Adolescent Smoking: Estimates, Correlates and Applications in Smoking Prevention Studies” *American Journal of Epidemiology*, Vol. 140, No. 11, pp. 1038-1050.

Pindyck, Robert S. and Daniel L. Rubinfeld (1998) *Econometric Models and Economic Forecasts*, 4th edition (Boston: Irwin, McGraw-Hill).

Prestine, Nona A. (1993) “Feeling the Ripples, Riding the Waves: Making an Essential School” in Joseph Murphy and Philip Hallinger eds. *Restructuring Schooling: Learning from Ongoing Efforts* (Newbury Park, CA: Corwin Press Inc.) pp. 32-62.

Raudenbush, Stephen, W. (1997) “Statistical Analysis and Optimal Design in Cluster Randomized Trials” *Psychological Methods*, Vol. 2, No. 2, pp. 173-185.

Raudenbush, Stephen W. (1994) “Random Effects Models” (Chapter 20) in Harris Cooper and Larry V. Hedges eds. *The Handbook of Research Synthesis* (New York: Russell Sage Foundation).

Rosenthal, Robert (1991) *Meta-Analytic Procedures for Social Research* (Newbury Park, California: Sage Publications).

SAS Institute Inc. (1989) “Chapter 44:The VARCOMP Procedure,” *SAS/STAT User’s Guide*, Volume 2, Version, 6 (Cary, NC: SAS Institute, pp. 1661-1667.

Slavin, Robert W., Nancy A. Madden, Alta H. Shaw, K. Lynne Mainzer and Mary C. Donnelly (1993) “Success for All: Three Case Studies of Comprehensive Restructuring of Urban Elementary Schools” in Joseph Murphy and Philip Hallinger eds. Restructuring Schooling: Learning from Ongoing Efforts (Newbury Park, CA: Corwin Press Inc.) pp. 84-113.

[i1] See, for example, Levy (1998), Blank (1997), and Burtless (1990).

[ii2] Murphy (1993) provides a brief review of these programs.

[iii3] See, for example, Levin (1993).

[iv4] See, for example, Prestine (1993).

[v5] See, for example, Barnett (1996).

[vi6] See, for example, Slavin et al. (1993).

[vii7] It might be possible to randomly assign students from a pool of volunteers to a “program school” that has

restructured (see Cave and Kagehiro, 1995). However, this is only possible under special circumstances. In addition, for some initiatives — such as a new curriculum component — it might be possible to randomly assign whole schools to a program or a control group. Raudenbush (1997) provides a theoretical discussion of the statistical properties of this "cluster random assignment" approach; and Bloom, Bos, and Lee (1998) provide an empirical analysis of these properties. In addition, Cook et al. (forthcoming) used the approach to evaluate Comer's Student Development Program in Prince George's County Maryland, and Cook et al. (1999) used it to evaluate Comer's program in Chicago. However, given the extensive "buy-in" required to restructure a school, it will not always be possible to use cluster random assignment in this context.

[viii8] Shadish (in preparation) provides a comprehensive review of the interrupted time-series literature. Campbell and Stanley (1966) and Cook and Campbell (1979) are perhaps the most widely cited sources on the topic.

[ix9] My thanks to Bob Granger of MDRC for raising these issues.

[x10] See Ham and Rock (1999).

[xi11] These data represent "moving cross-sections" for different third-grade cohorts each year. Hence, they represent different students in different years (with the exception of those who are held back and consequently retested).

[xii12] The parameters in Equation 1 are *identical* to their counterparts in Figure 1 when the number of students per year is constant. If the number of students varies over time, an appropriately weighted analysis of the annual means in Figure 1 will reproduce the parameter estimates for Equation 1.

[xiii13] Selection bias exists if the program did not cause the shift in student mix. However, selection does not create a problem of *bias* if the shift in student mix was caused by the program. It does, however, create a problem of *interpretation*. Without more data and/or further assumptions one cannot distinguish between: (1) the change in test scores due to the change in student mix caused by the program, or (2) the change in test scores due to the change in individual achievement caused by the program. In both cases, however, the change in test scores was caused by the program.

[xiv14] Idiosyncratic events and major changes in student mix should be distributed randomly across schools which operate independently of each other. Some of these changes will cause student achievement to be above the baseline trend, and others will cause student achievement to be below the baseline trend. Across a large number of schools, however, the average deviation from trend should be zero.

[xv15] We are basing our evaluation of the Accelerated Schools Project on the experiences of eight to ten schools from six to seven states that launched their programs between 1990-91 and 1993-94. These schools have consistent test data for a ten-year analysis period and were judged by staff from the national Accelerated Schools Project to have reached "mature" acceleration.

[xvi16] For the Accelerated Schools evaluation, which involved eight to ten school districts in six to seven states, we found it very difficult to obtain comparison school data. We thus are using a multiple program school design without comparison schools.

[xvii17] Ham and Rock (1999) describe the integrated quantitative and qualitative approach being used to evaluate Accelerated Schools.

[xviii18] Bloom (1995) illustrates how the minimum detectable effect of an estimator is a simple multiple of its standard error. This multiple depends on three factors: (1) the desired level of statistical significance, (2) the desired level of statistical power, and (3) whether a one-tail or two-tail test is being used.

[xix19] Effect size measures are especially popular for meta-analyses which pool impact estimates across different outcomes and metrics. For example, see Hedges and Olkin (1985) or Rosenthal (1991).

[xx20] For example, the mean value of five baseline years (-5 through -1) is -3.

[xxi21] For example, the sum of squared variation for five baseline years (-5 through -1) is 10.

[xxii22] The finding that precision declines as one forecasts further beyond the data used to estimate a model is well known in econometrics (e.g., see Pindyck and Rubinfeld, 1998, pp. 204-209).

[xxiii23] Throughout this paper I assume that the same test is used over time for a given school, although different versions of the same test may be used. This constraint is imposed on the design to avoid shifts in scores that occur when a new test is implemented. Linn, 1998, pp. 11-12, illustrates how scores fall precipitously when a new test is implemented and then rebound over time as teachers learn how to prepare students better (i.e., teach to the test). To evaluate the Accelerated Schools Project, we therefore only selected schools that used the same test for at least five years before and after launching their program.

[xxiv24] Equation 3 is a "random-effects" model. Such models are used in: (1) the econometrics literature on panel

data (see, for example, Greene, 1997, Chapter 14); (2) the statistics literature on analysis of variance (see, for example, Hays, 1973, Chapter 13); (3) the meta-analysis literature on pooling findings across studies (see, for example, Raudenbush, 1994), and (4) the literature on "hierarchical linear models" (see, for example, Bryk and Raudenbush, 1992, Chapter 2).

[xxv25] See Raudenbush (1997) for a discussion of the statistical issues involved.

[xxvi26] See Murray and Short (1995) Tables 2-8, pp. 685-691.

[xxvii27] See Murray et al. (1994), p. 1042.

[xxviii28] See Hannan et al. (1994), abstract, p. 88.

[xxix29] My thanks to Michelle Moser and Steve Caso for making these data available.

[xxx30] This represents a special case of aggregation where the value of each independent variable (the year indicator, t_i, and the dummy variables for each follow-up year, F_{0i}, F_{1i}, F_{2i}, F_{3I}, and F_{4i}) is the same for all members of the same aggregate unit (a cohort of students for a given year). Kmenta (1971), pp. 322-325, demonstrates that when this condition is met, the point estimates from an aggregate regression are identical to those from the model for individuals.

[xxxi31] To simplify the discussion, Equation A1 includes all baseline years but only one follow-up year. Equation 2 in the paper generalizes the model to include all follow-up years.

[xxxii32] Thus, the i^{th} student in year one is different from the i^{th} student in year two, and so on.

[xxxiii33] Equation B6 thus represents the minimum detectable effect size for a one-tail hypothesis test at the 0.05 significance level with 80 percent power.

[xxxiv34] See SAS Institute Inc. (1989).

[xxxv35] This assumes that random variation in tests scores for program schools and comparison schools are independent of each other.