Does Whole-School Reform Boost Student Performance? The Case of New York City

Robert Bifulco
William Duncombe
John Yinger

Abstract

Thousands of schools around the country have implemented whole-school reform programs to boost student performance. This paper uses quasi-experimental methods to estimate the impact of whole-school reform on students' reading performance in New York City, where various reform programs were adopted in dozens of troubled elementary schools in the mid-1990s. This paper complements studies based on random assignment by examining a broad-based reform effort and explicitly accounting for implementation quality. Two popular reform programs—the School Development Program and Success for All—were not found to significantly increase reading scores but might have been if they had been fully implemented. The More Effective Schools program was found to boost reading scores, but this effect seems to disappear when the program "trainers" leave the school. © *2005 by the Association for Public Policy Analysis and Management*

INTRODUCTION

Student performance in school districts with concentrated poverty, particularly large city districts, is far below student performance in other districts. Over the last decade, whole-school reform programs have been widely used to address this problem. This paper draws on the experience of New York City, where various whole-school reform programs were adopted in dozens of troubled schools in the mid-1990s. We explore the impact of these programs on students' reading performance.

Whole-school reform programs, which offer standardized sets of management and instructional prescriptions, stand out for two reasons. First, whole-school reform programs focus on the school as the unit of improvement, which distinguishes them from strategies that focus on system-wide policies and larger governing institutions. Second, these programs address, in a coordinated fashion, multiple aspects of school operations, including decisionmaking, resource allocation, classroom organization, curriculum and instruction, parental involvement, and student support. Traditional school-level interventions usually focus on one of these issues.[1]

[1] For a concise summary of several different models, see NWREL (1998).

Manuscript received July 2003; review complete December 2003; revised manuscript submitted April 2004; revision review complete June 2004; revision submitted July 2004; accepted July 2004.

Journal of Policy Analysis and Management, Vol. 24, No. 1, 47–72 (2005)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20069

Efforts to implement whole-school reform have been accelerating, particularly in urban schools that serve disadvantaged students. The Comprehensive School Reform Demonstration program (CSRD), enacted by Congress in 1997 and reauthorized in 2002 for $260 million, provides grants for schools to adopt "research-based" school-wide reform models. Moreover, in the spring of 1998, the New Jersey Supreme Court required hundreds of urban schools to implement (and the state to pay for) Success for All (SFA) (Goertz & Edwards, 1999). In addition, Memphis and Miami, along with New York City, have undertaken ambitious efforts to implement whole-school reform models.[2] As a result of efforts such as these, 24 different whole-school reform models had been adopted in over 8,300 schools nationwide by 1998 (Herman et al., 1999). Since then, SFA has been adopted in nearly 1,000 more schools and hundreds of schools have initiated whole-school reform efforts through CSRD.

The existing scholarly literature does not clearly reveal the effects of whole-school reform models on student academic achievement. Barnett (1996) reviews three of the most widely disseminated whole-school reform models: Accelerated Schools (AS), the School Development Program (SDP), and SFA. This early assessment concluded that "all three models can be implemented as described by their developer without substantial increases in per pupil school expenditures," but that the "evidence for the models' effects on educational outcomes for disadvantaged children is more ambiguous." A publication of the National Research Council concluded that whole-school reform designs have "achieved popularity in spite rather than because of strong evidence of effectiveness" (Ladd & Hansen, 1999, p. 153).

More recent contributions include evaluations of SDP in Prince George's County, Maryland (Cook et al., 1999), Chicago (Cook et al., 1998), and Detroit (Millsap et al., 2001), two of which randomly select treatment schools. Results from these studies are mixed, and suggest that SDP may not consistently result in improved student performance. Another study (Bloom et al., 2001) provides a multi-site, quasi-experimental evaluation of AS, also with mixed results. A quasi-experimental evaluation of the New York Networks for School Renewal Project (Schwartz, Stiefel, & Kim, 2004), which draws on the same data sets as this study, finds evidence of positive short-term program impacts. In 2001, the United States Department of Education funded six studies of whole-school reform models, the results of which are not yet available.

This article provides a quasi-experimental study of whole-school reform efforts in New York City in the mid-1990s. Several features of this study help to advance the emerging efforts to assess whole-school reform. Perhaps most importantly, the sites examined were part of large-scale efforts to implement whole-school reform in many schools, and thus provide evidence about the usefulness of whole-school reform as a broad school improvement strategy. The schools in the study were not identified as evaluation sites prior to model adoption, and thus reflect what is likely to happen in large-scale implementation efforts.

Because quasi-experimental approaches do not strive to control the implementation environment, are less expensive, and allow for the examination of many implementation sites, they provide an important complement to experimental studies. The primary challenges that face a quasi-experimental study are to make sure treat-

[2] Lacking clear evidence of program success based on its own evaluation, Memphis eventually cancelled its whole-school reform program. See Viadero (2001).

ment and comparison schools are really comparable and to account for the possibility that treatment schools are self selected (Cobb-Clark & Crossley, 2003). For example, if unobserved factors that influence a school's decision to adopt a whole-school reform also influence student performance in the school, then cross-sectional comparisons of adopting schools with non-adopting schools might provide biased estimates of model impacts. Our methodology, which is discussed in later sections, addresses these issues in detail.

Our analysis of program impacts focuses on two questions: (1) What is the cumulative impact of a whole-school reform model on student performance from first grade through third grade? (2) What portion of the one-year gain in student performance in the third, fourth, and fifth grades is attributable to one of these models? Several whole-school reform models focus on students in early elementary school, and the first question is designed to determine whether program impacts coincide with this focus. The second question is designed to determine whether these programs continue to boost student performance in later elementary school.

DATA

Treatment Sample

This study examines New York City elementary schools that adopted one of three whole-school reform models during the 1994–95, 1995–96, or 1996–97 school year. These models are the School Development Program (SDP), Success for All (SFA), and More Effective Schools (MES).[3] The schools in the top panel of Table 1 adopted a whole-school reform model in response to the New York State Education Department's (NYSED) registration review program. Under an initiative called Models of Excellence, NYSED facilitated and funded the adoption of whole-school reform models in the state's most troubled schools, called Schools Under Registration Review (SURR).[4] Adoption of a whole-school reform model was not required, however, and many SURRs did not adopt one. Our treatment group includes 24 SURR schools; 12 that adopted SDP, 9 that adopted MES, and 3 that adopted SFA.[5]

In addition, 2 of the 32 Community School Districts (CSD) in New York City undertook their own efforts to promote the adoption of whole-school reform. One of these implemented SDP in each of its schools in 1994–95. The other encouraged its elementary schools to adopt SFA, and 6 of them did so during the 1995–96 and 1996–97 school years. One school that independently adopted MES is also included in the sample. See the second panel of Table 1. In all, 47 schools adopted SDP, SFA or MES between the 1994–95 and 1996–97 school years.

[3] For brief descriptions of SDP and SFA see Barnett (1996) or NWREL (1998). For more complete descriptions see Comer, Haynes, & Joyner (1996) and Slavin et al. (1996). For a description of MES, see the Association for Effective Schools, Inc.'s Web site at http://www.mes.org.

[4] During the period examined by this study, a school was identified for registration review if it fell below any of the following criteria and showed a three-year pattern of decline on a criterion it failed to meet. The criteria, based on the state's Pupil Evaluation (PEP) tests, were 65 percent of students scoring above the state reference point (SRP) in third grade reading, 65 percent above the SRP in sixth grade reading, 85 percent above the SRP on eighth grade reading, 75 percent above the SRP in third grade math, and 75 percent above the SRP in sixth grade math.

[5] Several schools in the control group are also SURR schools. See Table 3.

Table 1. Whole-school reform model adopters included in the study sample.

	Model	Total Number of Adopters	Number Adopting in Fall 1994	Fall 1995	Fall 1996
SURR adopters	SDP	12	9	1	2
	MES	9	0	6	3
	SFA	3	2	0	1
	Total	24	11	7	6
Other adopters	SDP	16	16	0	0
	MES	1	0	0	1
	SFA	6	0	4	2
	Total	23	16	4	3
Total adopters	SDP	28	25	1	2
	MES	10	0	6	4
	SFA	9	2	4	3
	Total	47	27	11	9

SDP = School Development Program; MES = More Effective Schools; SFA = Success for All.

Comparison Group Selection

The selection of the comparison group is a critical issue in any quasi-experimental evaluation. As discussed in Heckman, LaLonde, and Smith (1999), a lack of comparability between treatment and comparison samples can lead to biased estimates of a program's impact, particularly if this impact itself varies with control variables, such as student characteristics. Our strategy is to select a large random sample of troubled schools to serve as the treatment group, and then to follow the advice of Cobb-Clark and Crossley (2003, p. 508) by trimming the sample so that the treatment and comparison schools have a similar propensity to adopt whole-school reform.

Stratified random sampling was used to select schools to serve as an initial comparison group. Beginning with all New York City elementary schools, we dropped schools from CSDs facing considerably different service delivery environments than the CSDs in which adopting schools are located.[6] Next, we created three sampling frames corresponding to the three years in which whole-school reform models were adopted. Each frame was limited to schools whose pre-adoption level of student performance fell below specified criteria,[7] and was split into quartiles based on student performance. An equal number of schools was randomly selected from each

[6] These districts serve few poverty students compared to districts with adopting schools, and in the typical year, do not have any schools with aggregate levels of performance that fall below the state criteria used to identify SURR schools (see footnote 4).

[7] We identify schools in which 55 percent or fewer students score above the SRP on the third grade PEP reading test *or* 70 percent or fewer students score above the SRP on the third grade PEP math test. A school had to meet this criterion in each of the three years before the relevant adoption year to be in the sampling frame. These criteria are similar to those defining SURRs (footnote 4) so they will yield a comparison group with a distribution of pre-adoption performance similar to that in the treatment schools.

quartile. The objective of this procedure was to produce a comparison group with a distribution of student performance that is reasonably close to the distribution in adopting schools. Overall, 28 schools were selected from the 1994–95 sampling frame, 12 from the 1995–96 frame, and 12 from the 1996–97 frame. Some schools were selected from more than one sampling frame. In addition, we dropped two schools because they said they had adopted a whole-school reform model in either 1997–98 or 1998–99. These steps yield 40 potential comparison schools.

We also identify a second, smaller set of comparison schools with observable characteristics that are matched to those of the treatment schools. More specifically, using our full sample, we first estimate logit models of the propensity that a school will adopt each whole-school reform program based on all observable school characteristics. In our analysis of each program, we then drop any comparison school with a propensity score below that of any treatment school. By comparing results for the full and trimmed samples, we can determine whether matching on school characteristics affects estimated program impacts.

The traditional approach to matching has been to use a sampling strategy that picks comparison schools with the same combinations of certain observable traits as the treatment schools. We have not followed that strategy for two reasons. First, this approach inevitably limits the dimensionality of any matching exercise, in the sense that only a few school characteristics can be brought into the sampling design. In contrast, propensity scores are based on all observable information (Heckman, LaLonde, & Smith, 1999). Second, sampling based on observable school characteristics can lead to differences in unobservable factors between the treatment and control schools, thereby magnifying the possibility of selection bias. To avoid these two problems, we draw a random sample of schools to form our comparison group and then use propensity scores to test the sensitivity of our results to closeness of the match between treatment and control groups.

Data Sources

Our main data come from individual student data files, called *Biofiles,* maintained by the New York City Board of Education (NYCBOE). We obtained data on all students who were in third grade in one of the sample schools during 1994–95, 1996–97, or 1998–99. These data include scores on NYCBOE's city-wide reading tests for each year the student took those exams. The NYCBOE did not administer the same test every year, so a simple comparison of test scores for different years might not yield an accurate picture of test score gains.[8] As shown in the next section, however, our estimation procedures do not require exact test score comparability. All test results are reported as Normal Curve Equivalents (NCE).[9] The NCE measure can be inter-

[8] Different tests measure different dimensions of reading performance and may use different norming procedures and samples. In 1994–95, the NYCBOE used *Degrees of Reading Power,* a test of reading comprehension developed by Touchstone Applied Science Associates. In 1995–96, NCYBOE switched to a reading test published by CTB/McGraw-Hill, and in 1998–99 began to use the reading component of the *TerraNova CAT.* These changes apply to both treatment and comparison schools, so they do not affect treatment/comparison differences in a given year, but they do imply that observed treatment/comparison differences in one year may not be strictly comparable to treatment/comparison differences in another year.

[9] The NCE is a test scoring metric developed to facilitate measurement of the effectiveness of Title I compensatory education programs. NCEs are normalized standard scores with a mean of 50 and a standard deviation of 21.06. Because NCEs form an equal-interval scale, they can meaningfully be aggregated and averaged (RMC Research Corporation [RMC], 1976), but NCE-based comparisons of different tests should be interpreted with caution.

preted as an equal-interval scale with a normal performance distribution. A gain of five NCEs represents the same amount of improvement at the extreme low (or high) end of the distribution as it does for average achievers.

There are 9,586 students in third grade in our sample schools in 1994–95. For those who remained in the New York City public school system and were not absent for, or exempted from, any tests, the data include test scores for each year from second through fifth grade. For the 9,932 students in third grade in 1996–97, the data include scores for third through fifth grade. For the 10,687 students in third grade in 1998–99, the data provide only third grade scores. The availability of test-score data is summarized in Table 2. This table also highlights the three years in which whole-school reforms were implemented in various schools (see Table 1). For each school in our sample, we observe students in each cohort.

Other annual information in the *Biofiles* includes the school the student attended and the student's grade, attendance information, and home zip code. In addition, the data set contains each student's date of birth, gender, ethnicity, home language, and school-lunch eligibility status during the spring of 1999.[10] These student-level files were linked with school-level data obtained from NYCBOE's Annual School Reports and from the NYSED's Basic Education Data System (BEDS).

The first four columns of Table 3 compare treatment and comparison schools along several dimensions potentially related to post-adoption performance. These figures are taken from the year prior to model adoption, or in the case of the comparison schools, from the year preceding the reference year used for the earliest sampling frame from which they were selected. These columns show that the student bodies of both treatment and comparison schools are almost entirely non-White, with a high percentage of students eligible for free lunch, although this percentage is slightly lower for SDP schools. MES schools are somewhat larger than the other treatment and comparison schools, with a majority of Hispanic students and a much higher share of students with limited English proficiency. The low levels of student performance before program adoption is demonstrated by the fact that a majority of students in all groups of schools fail to reach the statewide reference point (SRP) on the New York State Pupil Evaluation Program (PEP) tests in reading.[11]

Table 2. Available test scores in NYC data, by cohort.

Student Cohort	School Year					
		Years of Program Implementation				
	1993–94	**1994–95**	**1995–96**	**1996–97**	1997–98	1998–99
1994–95	2	**3**	**4**	**5**		
1996–97				**3**	4	5
1998–99						3

[10] Free-lunch eligibility indicators were imputed for 15.8 percent of the students in the study sample. Details on the imputation procedure used, and other aspects of the data assembly for this study, are available from the authors upon request.

[11] The PEP test and associated SRP, which is a minimum competency standard, were used until 1998–99 to identify students for remedial assistance. Impacts on this measure of school performance are presented in Bifulco and others (2002).

Table 3. Means (and standard deviations) for schools in the study sample.[a]

	Adopters						Comparison							
	SDP		MES		SFA		All		SDP Trimmed		MES Trimmed		SFA Trimmed	
Number of schools	28		10		9		40		26		14		18	
Number of SURR schools[b]	12		9		5		9		9		8		8	
Enrollment	753	(273)	1050**	(348)	886	(242)	749	(297)	721	(199)	847	(426)	780	(274)
% Asian	0.6	(0.9)	1.0	(1.0)	1.5	(1.4)	0.7	(1.4)	0.6	(0.8)	1.2	(2.1)	0.6*	(0.8)
% Black	67.4**	(28.5)	32.7**	(29.0)	60.2	(18.8)	51.0	(29.9)	64.6	(27.6)	39.7	(29.4)	57.9	(29.4)
% Hispanic	30.0**	(27.1)	64.8**	(29.2)	37.0	(17.4)	46.2	(28.5)	32.7	(25.8)	55.8	(26.5)	38.7	(27.1)
% White	1.8	(2.9)	1.4	(3.1)	0.9	(0.9)	1.8	(3.9)	1.9	(4.2)	3.2	(6.0)	2.6	(4.9)
% limited English proficient	13.6	(13.1)	32.8**	(23.2)	19.1	(13.3)	19.1	(14.8)	13.7	(12.5)	23.0	(14.9)	16.0	(12.7)
% eligible for free lunch	87.8**	(8.4)	93.7	(6.6)	94.1	(5.8)	91.9	(7.6)	91.5*	(8.0)	92.1	(6.9)	91.7	(8.3)
Average class-size	27.4	(2.5)	28.4	(3.6)	27.4	(3.1)	27.6	(2.6)	27.1	(2.6)	27.0	(3.3)	27.8	(2.7)
% teachers < 2 years experience	12.1	(7.1)	12.1	(4.5)	7.9	(5.7)	10.9	(7.8)	10.8	(7.1)	11.8	(7.3)	9.7	(5.4)
% teachers certified in field of assignment	79.5	(9.9)	76.3	(9.9)	90.1*	(6.5)	80.5	(12.4)	82.8	(12.0)	81.5	(8.0)	86.0	(8.7)
% above SRP on grade 3 PEP reading	49.0	(14.9)	44.1	(7.1)	46.9	(11.8)	46.5	(10.4)	47.0	(11.7)	43.8	(8.5)	45.6	(15.5)
% above SRP on grade 3 PEP math	72.1	(11.5)	76.7	(5.8)	80.7	(5.7)	76.3	(8.6)	75.4	(9.1)	74.8	(9.8)	78.6	(9.2)
Average Propensity Score (SDP)	0.130**	(0.10)					0.078	(0.08)	0.123	(0.10)				
Average Propensity Score (MES)			0.213**	(0.16)			0.023	(0.05)			0.053**	(0.08)		
Average Propensity Score (SFA)					0.125**	(0.14)	0.019	(0.03)					0.038*	(0.05)

[a] Reported averages and standard deviations are for the last year prior to program adoption, except for % of SRP on grade 3 PEP reading and math, which are averages for the three years immediately preceding model adoption. In the case of comparison schools, figures are from the year (or three years) preceding the reference year used to define the earliest sampling frame from which the school was selected.
[b] Counts all schools designated as a registration review school prior to 1997.
A * in the first three columns indicates significantly different than the comparison group mean (column 4) at the 0.10 significance level. In the last three columns, it indicates significantly different from the corresponding treatment group at the 0.10 level.
A ** indicates significantly different than the comparison group mean (column 4) at the 0.05 significance level.

Table 3 reveals that the three groups of treatment schools often differ significantly from the complete set of comparison schools in terms of enrollment, ethnic composition, poverty concentration, and share of students with limited English proficiency. As discussed earlier, differences of this type could lead to biased estimates of program impacts, and we use propensity scores to obtain a trimmed sample of comparison schools that is better matched to the treatment schools. The characteristics of the schools in our trimmed samples are presented in the last three columns of Table 3. None of the differences between the treatment and trimmed-sample comparison groups is significant at the 5 percent level.[12]

ECONOMETRIC METHODOLOGY

Production Function Framework

This study draws on the large literature concerning educational production functions, such as Ferguson and Ladd (1996). A general form for such a function is

$$Y_{ijT} = \alpha X_{ijT} + \beta W_{ijT} + \sum_{t=1}^{t=T-1} \lambda^{T-t} (\alpha X_{ijt} + \beta W_{ijt}) + \mu_i + \delta_j + \gamma_T + \varepsilon_{ijT}, \qquad (1)$$

where Y is a test score for student i in school j in year T and X is a set of explanatory variables, including both student and school characteristics.[13] The variable of interest in this study is W, which indicates that a school has implemented whole-school reform. (This variable is discussed in more detail below.) The coefficient of this variable, β, is our measure of program impact. The effect of explanatory variables from previous years carries over to year T but degrades at a rate given by $(1-\lambda)$. This form also contains a year fixed effect, γ, and time-invariant fixed effects for the student, μ, and the school, δ. The final term represents random error.

PROCEDURES TO ELIMINATE SELECTION BIAS

The main challenge we face in estimating (1) with non-experimental data is that the estimate of β may be biased if unobserved characteristics of a school, which include δ and may include lagged values of the school-level Xs, are correlated with the decision to adopt whole-school reform. This type of bias is often called self-selection bias.[14]

[12] Significance tests reported in Table 3 were conducted using a school as the unit of analysis. We also used our sample of third graders in 1995 to test for statistically significant differences between treatments and their matched comparison groups on five student level variables: Black, Hispanic, eligibility for ESL services, free-lunch eligibility, and reading test score. After adjusting for clustering within schools, the only statistically significant difference is that MES students were significantly more likely to be eligible for ESL services than students in their matched control schools.

[13] Because our data set combines school-level and individual-level variables, standard procedures will understate the standard error of β (and of other coefficients). To avoid this problem, we calculate standard errors using the "cluster" option in Stata (Stata Corporation, 2003), which is based on a generalization of the Huber/White "sandwich" estimator of variance (see Greene, 2003, pp. 520–521). An alternative way to avoid understating standard errors is to focus on school-wide average test scores (see Bloom et al., 2001). This type of analysis is also presented below.

[14] In addition, students with certain unobserved characteristics (included in μ) or with certain past experiences (included in the lagged values of the individual-level X's), might have a tendency to move to schools in which whole-school reform has been implemented. This type of correlation, along with a correlation between these moving decisions and student performance, might also lead to biased estimates of β. See footnote 16.

Several methods have been developed to eliminate this selection bias. First, suppose that the individual and school fixed effects equal zero. In this case, the test score in a previous year can be used to account for the effect of explanatory variables from previous years, even if they are not observed. Specifically, setting μ and δ equal to zero and subtracting λ times Y_{ijT-1} from Y_{ijT} yields a standard "value-added" formulation of a production function:

$$Y_{ijT} = \alpha X_{ijT} + \beta W_{ijT} + \lambda Y_{ijT-1} + (\gamma_T - \lambda\gamma_{T-1}) + (\varepsilon_{ijT} - \lambda\varepsilon_{ijT-1}). \quad (2)$$

Note that the expression containing γ serves as a constant term; it captures changes over time in the nature of the test or in the average score of participating students.

Second, suppose that λ equals zero, that is, that the impact of the Xs in previous years does not carry over. In this case, we can subtract Y_{ijT-1} from Y_{ijT} to obtain a "difference" formulation of a production function, which eliminates μ and δ. This form is:

$$Y_{ijT} - Y_{ijT-1} = \alpha(X_{ijT} - X_{ijT-1}) + \beta(W_{ijT} - W_{ijT-1}) + (\gamma_T - \gamma_{T-1}) + (\varepsilon_{ijT} - \varepsilon_{ijT-1}), \quad (3)$$

where, as before, the expression containing γ serves as the constant term.

Both of these formulations require two years of data, but the value-added version, equation (2), does not require two years of data for the explanatory variables. The value-added approach can be applied to two observations after implementation or to one observation before and one after implementation because the impact of explanatory variables in previous years, including W, is summarized in the lagged dependent variable. This is not true for equation (3). When W is differenced, it equals zero if the program is in place in both years and therefore drops out of the analysis. As a result, the difference approach requires at least one observation before program implementation and one observation after. With this type of data, equation (3) compares the change in test scores in schools that implement the program with the change in test scores in schools that did not do so, and therefore provides a "difference-in-difference" estimator.

Neither equation (2) nor equation (3) is satisfactory by itself, because neither one accounts for both fixed effects (individual and school) and carryover effects through the X variables.[15] After all, either of these factors could result in biased estimates of β. One way to account for both of these factors is to combine the steps that lead to equations (2) and (3), that is, to difference a value-added model without first setting the fixed effects equal to zero. This approach requires three years of data. Unfortunately, however, this approach, like equation (3), requires pre-implementation information and, in fact, requires two years of data before program implementation and one year after. Only a small subset of our data meets these requirements, and then is only suitable for examining the impact of whole-school reform on the change in test scores in fifth grade.

Another approach, the one we use, is to estimate equation (2) using an instrumental-variables (IV) technique that accounts for the potential impact of unob-

[15] More technically, subtracting λ times equation (1) for $T-1$ from equation (1) results in an equation (2) with two new terms on the right side, namely, $\mu_i(1-\lambda)$ and $\delta_i(1-\lambda)$. Subtracting equation (1) for $T-1$ from equation (1) results in an equation (3) with the following new terms on the right side: $S_{T-1}(\lambda - 1) + \lambda(\alpha X_{ijT-1} + \beta W_{ijT-1})$, where S_t is the sum in equation (1) for year t.

served school characteristics, δ, on the decision to adopt whole-school reform, W.[16] To identify instruments for W, we hypothesize that a school will be more likely to adopt a given model if other schools in the same district (that is, CSD) have already done so. The presence of other adopting schools in the same district makes it more likely that a school will have information on a model, thereby reducing search costs; provides opportunities for jointly purchased training, potentially reducing implementation costs; and might enhance the perceived professional advantages of adoption.

Because schools in the same district may draw their students from similar populations and use a similar, district-level hiring process, unobserved characteristics of students and teachers in schools from the same district might be correlated. This implies that the number of schools in the district that have adopted a whole-school reform model may not be an exogenous source of variation in a school's decision to adopt. If, however, the decision of other schools in the district is driven primarily by observed characteristics of those schools, then these observed characteristics provide suitable instruments. If a school is influenced by the other schools in its district, then the observed characteristics of those other schools, which influence their own propensity to adopt a whole-school reform model, provide predictors of the initial school's propensity to adopt. Moreover, observed characteristics of other schools are unlikely to have any direct influence on student performance in the initial school. Thus, we use the average characteristics of other schools in the same district to identify exogenous variation in W.

Linking Methods to Research Questions

With our data, a value-added formulation, equation (2), with W treated as endogenous is ideal for examining the second research question defined earlier, namely, the impact of whole-school reform on a student's progress in grades 3, 4, and 5, each estimated separately. This approach is not possible, however, for our first research question, namely, the cumulative impact of whole-school reform in the early elementary years, because it requires a pre-grade-one test score. Such a score does not exist in our data.[17]

Fortunately, however, we can answer our first research question using our IV technique without using a value-added specification. This technique directly addresses the most likely source of bias in β, namely the correlation between δ and program implementation. In addition, this technique is an appropriate method for dealing with the potential bias in β that arises if the lagged school-level Xs in equation (1) are correlated with W. The exclusion of the lagged dependent variable lowers the explanatory power of the regressions and is therefore likely to raise the stan-

[16] This approach does not address the potential correlation between unobserved individual characteristics, μ, and W; that is, it does not eliminate biases that might arise if parental choices about where to live and send their children to school are influenced by whole-school reform decisions. To put it another way, our approach eliminates bias associated with the whole-school reform adoption decision itself, but not from parents' behavioral responses to this decision. We suspect, however, that few parents are even aware of decisions about whole-school reform and that fewer still respond to them. Schwartz, Stiefel, and Kim make a similar argument (2004, p. 510): "Students or their parents are unlikely to have known about this reform before enrolling (or even after enrolling)."

[17] State-administered tests typically are not given in kindergarten or first grade because test-taking skills generally have not been developed by that age. The lack of these tests is therefore an inherent constraint facing research on whole-school reform.

dard errors of the coefficients, but this variable is not required to eliminate self-selection bias in β.[18]

We took several additional steps to verify the validity of our IV strategy. First, we used over-identification tests to confirm that the instruments used in each regression are not correlated with unobserved factors that influence student performance (Wooldridge, 2002). Second, we used procedures described by Bound, Jaeger, and Baker (1995) to verify that our instruments explain a significant share of the variation in treatment status. Finally, we used the subset of students for whom two or more pre-exposure measures of performance are available to compare our IV results with those obtained using a value-added, difference-in-differences estimator. As shown earlier, this estimator accounts for the unobservables in equation (1).[19]

The link between our data and our research questions is explained in detail in Table 4. The rows of this table refer to the substantive research questions we plan to address, namely, the cumulative impact of whole-school reform in grades 1 to 3, and the value-added impact of whole-school reform in grades 3, 4, and 5. The first two columns indicate the combinations of a student cohort and a year of implementation that will be used to answer each question. For example, the 1994–95 cohort of students cannot be used to help answer the first question, because, as shown in Table 2, the students in that cohort were in grade 3, 4, or 5 when whole-school reform was implemented in their school. In the 1996–97 cohort, on the other hand, students in schools that adopted whole-school reform in 1994–95 experienced whole-school reform starting in the first grade, so they have spent their entire early elementary years in a whole-schools reform school by the time we observe their third grade scores in 1996–97.

The last two columns of Table 4 indicate the number of treatment schools (for each reform model) and the number of observations (that is, students) in treatment schools available to answer each substantive research question. For example, our answer to the first research question for SFA will be based on 9 treatment schools and 6,570 observations, 885 of which are in treatment schools.

Missing Test Scores and Student Mobility

Across the three cohorts, approximately 34.2 percent of students are missing at least one reading test score.[20] The students with missing test scores are more likely

[18] As an anonymous reviewer pointed out, an additional source of potential bias is created by our selection of students who have reached third grade at given points in time. Whole-school reform models can influence whether or not a student reaches third grade with his or her original cohort. SFA, in particular, discourages student retention as a matter of policy. As a result, low-performing students might be more likely to reach third grade with their cohort in treatment group schools than in comparison group schools, which would create downward bias in the estimates of cumulative third grade treatment impacts. At the same time, third graders in comparison group schools might be more likely to include low-performers retained from preceding cohorts, which could create upward bias in estimated treatment impacts. The net effect of this bias could be positive or negative. Students in our sample from SFA schools in particular are younger on average and less likely to be overage for their grade than the comparison group students. To address this potential source of bias, we include school level measures of the percent of overage students as a control variable in our regressions. Inclusion of this control increases estimates of cumulative third grade impacts for SFA, but only slightly, and as reported in Table 5, estimates remain statistically insignificant.

[19] This analysis is presented in Bifulco (2002). That article also provides additional discussion of the conditions required for the difference-in-difference and IV estimators to provide unbiased and/or consistent impact estimates.

[20] The percentage of students missing the test scores needed for the specific analyses presented here is less than 34.2 percent and varies by cohort and school year.

Table 4. Sources of data for key questions.

	Sources of Data		Data Description	
	Cohorts of Students	Implementation Years	Number of Treatment Schools (Model)	Number of Observations in Treatment [Comparison] Schools
Question 1: Cumulative impact				
Grades 1 through 3	1996–97	1994–95	28 (SDP)	3,253 [5,685]
	1998–99	1994–95	10 (MES)	855 [5,685]
		1995–96	9 (SFA)	885 [5,685]
		1996–97		
Question 2: Value-added impact				
Grade 3	1994–95	1994–95	25 (SDP)	1,827 [2,771]
Grade 4	1994–95	1994–95	28 (SDP)	3,483 [5,933]
		1995–96	10 (MES)	1,511 [5,933]
	1996–97	1994–95	9 (SFA)	1,208 [5,933]
		1995–96		
		1996–97		
Grade 5	1994–95	1995–95	28 (SDP)	3,156 [5,105]
		1995–96	10 (MES)	1,794 [5,105]
		1996–97	9 (SFA)	1,288 [5,105]
	1996–97	1994–95		
		1995–96		
		1996–97		

than other students to be male, to be Asian or Hispanic, to be eligible for free lunch, to speak a language other than English at home, to be eligible for ESL services, and to have changed schools.

Whether or not missing test scores bias estimates of whole-school reform model impacts depends on the answers to two questions. The first question is whether or not a student's enrollment in a school that has adopted a whole-school reform model is independently related to that student's having a test score reported. For most of the analyses we conduct, this is not the case. Nonetheless, for some cohorts, in some years enrollment in a whole-school reform model does show a statistically significant influence on the probability of observing a complete set of test scores, even after controlling for other student characteristics. The second question is whether or not students with missing test scores would, if they were tested, tend to have different scores or score gains than otherwise similar students for whom we do observe test scores. This question cannot be answered with our data, so an affirmative answer cannot be ruled out.

This missing test score issue is compounded by the fact that students in one of our sample schools in third grade might have moved to a school outside our sample dur-

ing or prior to the year being examined. For example, 22.5 percent of the cohort in third grade in 1994–95 moved to a school not included in the study sample by fifth grade. Although the data set allows us to follow these students into schools outside the study sample, the schools into which these students have moved might be substantially different in terms of student-body characteristics, resources, and efficiency than the schools that have adopted whole-school reform. Comparison of student performance in whole-school reform schools with the performance of students in markedly different schools can produce misleading estimates of the impacts of whole-school reform. Thus, the primary analyses in this study are conducted using only students who have remained in one of the treatment and/or comparison group schools.[21]

In sum, excluding students who have missing test scores or have moved to schools outside the study sample may bias estimates of model impacts. To address this potential bias, we employ a Heckman two-step selection-correction procedure (Heckman, 1979; Greene, 2003).

The Role of Implementation

We define treatment as the decision to adopt whole-school reform. If the decision to adopt does not have a large impact on student performance, our basic approach cannot distinguish a failure of the model's prescriptions to improve student performance from a failure of treatment schools to consistently implement those prescriptions. To help make this distinction, we also estimate whether the impact of the adoption decision depends on the quality of implementation, based on information provided by the program developers.[22]

The impact of whole-school reform also might depend upon a school's experience with a particular whole-school reform model. Similarly, student mobility implies that not all students in a treatment school have been exposed to whole-school reform for the same number of years, and a reform program's impact may increase with the length of time a student has been exposed to it. To account for these possibilities, we also ask whether program impacts vary with the length of time a school has been implementing a particular reform model or with the number of years a student is exposed to a reform program.

RESULTS

Main Results

Our main results are presented in Tables 5 and 6.[23] These tables present results for each research question for each whole-school reform program using three different

[21] A small number of students in each cohort moved from one sample school to another school that is also in the sample. These students are included in the primary analyses. Our approach contrasts with that of Schwartz, Stiefel, and Kim (2004), who retain in their sample all students who remained in any New York City school.

[22] The distinction between schools that decide to adopt a whole-school reform model and schools that are able to implement that model's prescriptions is analogous to the distinction between individuals assigned to a treatment group and those who actually receive the treatment in randomized experiments (Rouse, 1998). For further discussion of this distinction in the case of whole-school reform, see Bifulco et al. (2002).

[23] These tables, along with the others in the text, focus on results for the whole-school reform variables. The full list of control variables is provided in Appendix Table A1, which presents full results for the fifth-grade, value-added regressions. Results for any other regression in this paper are available from the authors upon request.

methodologies. All three methodologies make use of a Heckman (1979) selection correction to account for missing test score information and student mobility.[24] The first methodology is ordinary least squares (OLS) with the full sample. The second methodology, designed to address the comparability problem, is OLS with the trimmed sample. The third methodology, designed to address the selection problem, is instrumental variables (IV) with the full sample.[25] We attempted to implement a fourth methodology designed to address both problems simultaneously, namely, IV with the trimmed sample. As it turns out, however, the limited variation in school characteristics in the trimmed sample makes it impossible in most cases to identify instruments that help to explain the adoption decision and pass an exogeneity test. As a result, we must be cautious in interpreting cases in which the second and third methodologies yield different results.

Table 5 presents results for our first question, namely, the cumulative impact of whole-school reform in grades one through three. For SDP (the first panel) and SFA

Table 5. Estimates of the cumulative impact of whole-school reform through grade 3, using alternative samples and estimators.[a]

	OLS	OLS	IV
	Full Sample	Trimmed Sample	Full Sample
SDP	1.079	0.763	1.914
	[1.422]	[1.444]	[2.484]
Uncensored students obs.	10,774	8,311	10,774
Censored obs.	8,938	6,963	8,938
Treatment schools	28	28	28
Comparison schools	40	26	40
MES	2.850**	3.600**	14.179**
	[1.431]	[1.501]	[6.815]
Uncensored students obs.	8,115	4,222	8,115
Censored obs.	6,540	3,067	6,540
Treatment schools	10	10	10
Comparison schools	40	14	40
SFA	1.474	1.143	3.295
	[1.298]	[1.937]	[4.729]
Uncensored students obs.	8,080	4,576	8,080
Censored obs.	6,570	3,602	6,570
Treatment schools	9	9	9
Comparison schools	40	18	40

[a] Estimates are each drawn from separate regressions controlling for several student and school characteristics, and using Heckman correction procedure. See Appendix. Figures in brackets are robust standard errors. A ** indicates a significance at the 0.05 level.

[24] To be specific, we estimate the probability that a student will have all test information and then insert the resulting selection-correction term into our student-performance regression. Results for this equation, which is estimated with probit analysis, are presented in Table A1. The selection-correction term is statistically significant in most of our OLS regressions, but usually is not significant with our IV procedure.

[25] All our IV estimates use the weighting procedure in Heckman, LaLonde, and Smith (1999, p. 1987), which is needed to ensure consistency with a choice-based sample.

(the third panel) all three methods yield similar results. For both programs, the estimated coefficient is slightly above 1.0 and statistically insignificant, regardless of which methodology is used. In contrast, the results for MES (the second panel) are positive and significant for all three methodologies. The NCE scale is designed to have a standard deviation (SD) of 21.06. Consequently, the point estimate using IV (the third column) implies that three years of exposure to MES raises the average student's test score by 68.7 percent of an SD, which is a large impact. The point estimate for the trimmed sample (the second column) is smaller, 17.1 percent of an SD. Because we cannot identify a set of valid instruments using the trimmed sample, we cannot determine which of these estimates better approximates an estimate obtained with a methodology that accounts for both comparability and selection.

The results for question two, concerning value-added impacts, are presented in Table 6. We find no significant impact for SDP for any grade or any methodology. In the case of MES, we find a large, positive, and significant impact in grade 4 with the IV methodology. This result does not appear with the trimmed sample, however, so we cannot rule out the possibility that this result reflects a lack of comparability between the treatment and comparison schools. Moreover, we were unable to resolve this issue by applying IV to the trimmed sample, because we were unable to identify a set of valid instruments in this case. Turning to SFA, we find negative and significant results in fifth grade using OLS on both the full and trimmed sample. The IV approach yields an even larger negative impact, but it is not statistically significant.[26] These results suggest that SFA's strong emphasis on reading in the early grades might attract attention away from reading in fifth grade; given the insignificance in the IV result, however, this conclusion should not be taken as definitive. Overall, there is no clear evidence that any of these programs have positive impacts on reading improvements in third, fourth, or fifth grades.

We also implemented several variants of our basic methodology.[27] First, for a selected subsample of our data, we are able to estimate fifth-grade value-added impacts using the difference-in-difference approach. This approach, which is possible only for MES and SFA, yields results similar to those in the last column of Table 6.[28] Second, we were able to correct for possible measurement error in the lagged test score in the fifth-grade value-added regressions for all three programs. This type of measurement error can lead to a correlation between the lagged test score and the error term in a value-added specification, and hence might result in biased estimates. To deal with this potential problem, we use an IV procedure, with test score lagged two years as the instrument. Again, the results are similar to those in the last column of Table 6. Third, we examine the impact of our selection correction for missing test scores by conducting, for each cell in Tables 5 and 6, two alternative analyses, one in which movers are included and one without the selection correction. These results are also approximately the same as those we present.

As pointed out earlier, the impact of whole-school reform may not be the same for all types of students. To determine whether model impacts depend on ethnicity,

[26] We were able to identify a set of valid instruments for the trimmed sample, and an IV estimate for the trimmed sample yields a statistically significant negative effect with about the same magnitude as the IV estimate for the full sample. This gives further support to the conclusion that SFA has a negative impact in fifth grade.

[27] Detailed results for these alternative approaches (and others discussed below) are available from the authors upon request.

[28] As in Table 5, the results are also similar whether or not an IV procedure is included. The impact of differencing on impact estimates with these data is explored in more detail in Bifulco (2002).

Table 6. Estimates of the value-added impact of whole-school reform, grades 3, 4, and 5, using alternative samples and specifications.[a]

	Grade 3			Grade 4			Grade 5		
	OLS Full Sample	OLS Trimmed Sample	IV Full Sample	OLS Full Sample	OLS Trimmed Sample	IV Full Sample	OLS Full Sample	OLS Trimmed Sample	IV Full Sample
SDP	1.976	2.360	0.480	-0.833	-0.877	2.948	-0.119	-0.224	-0.302
S.E.	[1.541]	[1.598]	[2.753]	[0.627]	[0.632]	[2.070]	[0.755]	[0.803]	[1.519]
Uncensored obs.	6,612	5,173	6,612	10,943	8,415	10,943	9,189	7,035	9,189
Censored obs.	4,598	3,838	4,598	9,416	7,432	9,416	8,261	6,406	8,261
MES				0.581	-0.166	5.929**	-0.008	-0.04	0.869
S.E.				[0.871]	[1.126]	[2.620]	[0.647]	[0.747]	[2.610]
Uncensored obs.				9,033	4,895	9,033	7,857	4,378	7,857
Censored obs.				7,444	3,852	7,444	6,899	3,776	6,899
SFA				0.211	0.566	-0.729	-2.315**	-2.078**	-3.269
S.E.				[0.886]	[0.824]	[2.801]	[0.641]	[0.554]	[3.327]
Uncensored obs.				8,484	4,920	8,484	7,150	4,206	7,150
Censored obs.				7,141	4,221	7,141	6,393	3,807	6,393

[a] Estimates are each drawn from separate regressions controlling for several student and school characteristics, and using Heckman selection correction. See Appendix. Figures in brackets are robust standard errors. A ** indicates a significance at the 0.05 level.

poverty, or English proficiency, we re-estimate each full-sample IV model with three interaction terms, namely, the program adoption variable interacted with each of these student characteristics. The only result that is statistically significant (and then at only the 10 percent level) appears in the MES equation for cumulative program impacts in grades 1 through 3. Specifically, we find that the positive impact of MES in these grades does not arise for students with limited English proficiency.

School-Level Analyses

As in other studies of how organizations affect individuals, one might ask whether schools or students are the appropriate unit of analysis. A common approach in studying organizational effects, and the one we have taken, is to use individuals (students) as the unit of analysis and to adjust standard error estimates for clustering within organizations (schools).[29] This approach is consistent with our question about the impact of the treatment on an individual level learning process. Individual level analyses also provide other advantages: Parameter estimates are less susceptible to being distorted by organizations with a small number of observations (Raudenbush & Bryk, 2002); corrections for missing test score information (an important potential source of bias) can be made, and analysis of treatment effects across types of students can be conducted.

Nonetheless, those who believe that schools are the appropriate unit of analysis will suspect that our student level analyses overstate the precision of our estimates. One might ask, then, whether the statistically significant findings in Tables 5 and 6 would remain significant in school level analyses. To investigate this topic we used our student samples to calculate school-level means for each of the variables in our student level model. We then ran each of our regressions using mean reading scores as the dependent variable and the mean values of each of the explanatory variables used in the student level models. The results of these analyses, for those cases in which the student level analyses found statistically significant impacts, are presented in Table 7.[30]

The point estimates in Table 7 have a different interpretation than the point estimates from the student level analyses reported in Tables 5 and 6. While both sets of estimates represent the average impact of whole school reform across adopting schools, the student level estimates represent student-weighted averages while the school level estimates represent unweighted averages. While both of these parameters are potentially of interest, if we are concerned with the impact of broad based efforts to disseminate whole-school reform models on students, the student-weighted averages might be more relevant.

The top panel of Table 7 presents estimates of the cumulative impact of MES on third grade reading achievement. These estimates have the same sign and similar magnitudes as the corresponding estimates in Table 5. However, the standard errors are larger here, and as a result the point estimates are statistically significant only for the OLS estimates using the trimmed sample, and even then, only at the 0.10

[29] See Raudenbush and Bryk (2002) (pp. 99–117) and Wooldridge (2002, pp. 328–331) for discussions. We use Huber-White robust standard errors to allow for clustering within schools. An alternative is to estimate coefficients and standard errors using maximum likelihood estimators as done in studies that use hierarchical linear modeling (HLM). An advantage of the approach we have used is that Huber-White robust standard errors place fewer restrictions on the variance of the error terms than HLM estimates.

[30] These school level analyses were conducted for all the cases presented in Tables 5 and 6. In all the cases not reported, impact estimates were statistically indistinguishable from 0.

Table 7. School-level analysis of cumulative and value-added impacts of whole-school reform.[a]

	OLS Full Sample	Grade 3 Cumulative OLS Trimmed Sample	IV Full Sample
MES	2.118	3.912*	15.074
S.E.	[2.164]	[2.108]	[18.239]
Treatment schools	10	10	10
Comparison schools	40	14	40
	OLS Full Sample	Grade 5 Value-Added OLS Trimmed Sample	IV Full Sample
SFA	−1.971**	−1.394*	−5.626
S.E.	[0.677]	[0.727]	[5.198]
Treatment schools	9	9	9
Comparison schools	40	18	40

[a] Estimates are each drawn from separate school level regressions. Other covariates include percent female students, percent Hispanic students, percent free-lunch eligible, percent of ESL students, log of enrollment, percent of teachers with less than 2 years experience, percent of teachers certified in their field, average class-size, a registration review indicator and a year indicator. For the value-added models the average lagged test score and the average lagged test score interacted with the year indicator are also included. For each analysis estimates are based on data pooled from two separate years, therefore, robust standard errors are used to account for any clustering within schools.

level. The estimated, value-added impact of Success for All in fifth grade remains negative and statistically significant when the full sample is used (see bottom panel of Table 7). When the trimmed sample is used the estimated impact is statistically significant only at the 0.10 level. Thus, the school level analyses suggest a more conservative conclusion that adoption of MES or SFA, like adoption of SDP, has no consistently discernible impacts on student achievement, positive or negative.[31]

Do Impacts Depend on Program Implementation?

The results above indicate that SDP and SFA did not have consistent, positive impacts on student performance. Because these estimates focus on the impact of the decision to adopt one of these models, it remains unclear whether the lack of positive impacts is due to a failure of the model's prescriptions to improve student performance or from a failure of treatment schools to consistently implement those prescriptions.

To study the impact of implementation for SDP and SFA, we collected extensive implementation data from the program developers. These data were used to develop measures of overall implementation quality at adopting schools, and to distinguish cases in which implementation was relatively successful from cases in

[31] This conclusion is not due to a lack of statistical power in these analyses. Although, many of the IV estimates lack adequate statistical power, the estimates of the cumulative impact of MES in third grade are the least precise of all the OLS estimates. The standard errors on these estimates indicate these analyses had a good chance (80 percent) of finding statistically significant results if the true impact were as small as +5.4 (for full sample, +5.2 for trimmed sample), which on the NCE scale represents approximately 0.26 standard deviations.

which implementation was less successful.[32] We could not obtain comparable information for MES, and these data did not cover all SDP schools. The implementation measure for SDP is an average of the program developer's ratings for school planning and management team effectiveness, mental health team effectiveness, parent team effectiveness, and comprehensive school plan effectiveness (Emmons, 1999). In the case of SFA, the implementation measure averages the program developer's estimates of success in assessment and regrouping, tutoring for reading, staff development and support, early learning, and curriculum.

To determine if program impacts depend on the degree of implementation, we interact the treatment variable with the quality of implementation variable, expressed as a deviation from the sample mean for schools in a given year.[33] A positive sign for this interaction term indicates that treatment impact increases with implementation quality. We only present OLS estimates of this impact because our instrumental variable strategy is not appropriate for this analysis. First, the SDP schools for which we have implementation ratings are all from the same district, undermining the usefulness of our instruments, which are based on district characteristics. Moreover, we do not have any new instruments to deal with the potential endogeneity of implementation quality. As argued above, our instruments have a clear conceptual link to the decision to adopt a whole-school reform program, but they do not have a strong conceptual connection to implementation quality. After all, the factors determining which schools adopt a whole-school reform model are not necessarily the same as those determining which schools successfully implement that model.

Our results are presented in Table 8. We find that cumulative program impacts increase with implementation quality. This result is significant at the 5 percent level for SFA and at the 10 percent level for SDP. Although the content of the implementation indexes is somewhat different for the two programs, the point estimate is virtually the same: A one-point increase in either index boosts the cumulative impact by about 16.6 percent of the SD in the test scores (3.5/21.06). This result does not imply, however, that the impacts of these models would have been large if they had been implemented well, because the average implementation rating was quite high already. Out of maximum of 4.0, the average ratings for SDP are 3.412 in 1997 and 3.552 in 1999. The average SFA ratings for the same two years, which have a maximum of 5.0, are 3.101 and 4.000, respectively. Thus, bringing all SFA schools up to the maximum implementation rating, which obviously would be difficult to accomplish, would boost test scores by about one-third of an SD in 1997 and one-fifth of a SD in 1999.[34] A comparable improvement in implementation for SDP would have a smaller impact, less than one-seventh of an SD.

[32] These ratings come from unpublished surveys supplied to us by the program developers. More information on the implementation ratings for SDP and SFA and on many other features of program implementation can be found in Bifulco and others (2002).

[33] This formulation implies that the estimated impact of W alone is still an estimate of the average impact of that model. This estimate differs from the estimates in Tables 5 and 6, however, because it is based on a slightly different sample (implementation data are not available for all SDP schools) and because the implementation rating is expressed as a difference from the school mean not from the student mean.

[34] In 1997, raising all districts to the maximum score would raise the average from 3.101 to 5.000, a change of 1.899. Multiplying this change by the point estimate, 3.499, yields an increase in the average test score of 6.645. Finally, adding the coefficient of W and dividing the result by the test-score SD, 21.06 yields the result in the text. A comparable calculation leads to the other results in the text and also indicates that the impact of SFA on fifth-grade value-added is negative even with perfect implementation (see the last column of Table 7). Note, however, that the SFA implementation ratings were used to provide feedback to the school and therefore served both a motivational and an evaluative function. The program developers told us that implementation ratings might have been inflated somewhat in order to maintain the motivation of teachers and staff.

Table 8. Variation in the estimated impact of whole-school reform by quality of model implementation.

	Value-Added Impacts		
	Through Grade 3 OLS	Grade 4 OLS	Grade 5 OLS
SDP	0.868	–0.893	–0.149
	[1.834]	[0.654]	[0.957]
SDP*implementation rating	3.574*	0.217	1.825*
	[1.996]	[0.733]	[1.089]
SFA	0.915	0.867	–2.249**
	[1.096]	[1.147]	[0.627]
SFA*implementation rating	3.499**	0.548	0.051
	[1.230]	[2.076]	[0.930]

Estimates are drawn from separate regressions for each program controlling for several student and school characteristics. See Appendix. SDP estimates computing using only SDP schools with implementation ratings, and SFA results computed only using student outcome measures from those years that we have SFA implementation measures (1997–1999), otherwise samples are the same as in Table 6. Figures in brackets are robust standard errors; * = significant at 0.10 level; ** = significant at 0.05 level.

Because our implementation measures are exploratory and because these regressions are not estimated with either a trimmed sample or an IV procedure, these results are no more than suggestive. With this caution in mind, they indicate that implementation may matter and that these two programs may not have substantial benefits unless implementation is complete. This finding reinforces the value of quasi-experimental studies, which are more likely than experimental studies to observe schools with a range of implementation ratings. It may also explain why SFA and SDP have significant impacts in some small studies, which can carefully control implementation, but not in larger venues like New York City, where implementation was difficult to control.

An additional feature of program implementation is that the impact of a whole-school reform program may change as a school gains experience implementing it or as students are exposed to it for a greater length of time. An exploration of these two effects for the cumulative impacts of MES in grades 1 through 3 yields some intriguing results. Specifically, we find that program impacts are greatest when schools have fewer than two years of experience with MES. Moreover, the impact of MES is positive but insignificant for students exposed to MES for one or two years, that this impact grows with every year of exposure, and that it is significant for students exposed for three or four years.[35]

One possible interpretation of these results is suggested by a key feature of MES implementation in the New York City schools. Specifically, the program developers sent trainers into the schools that adopted MES. In addition to providing initial instruction on the model's precepts, these trainers visited schools weekly to assist the school planning team in administering school surveys, analyzing survey and performance data, and conducting school planning processes. These trainers were present in the MES schools during 1995–96 and 1996–97, but were not present in

[35] Complete results on these two effects for MES and comparable results for SDP and SFA (and for value-added impacts) are available from the authors upon request.

any of the MES schools in 1997–98 or 1998–99.[36] This timing coincides with our measures of school experience. When schools are observed with fewer than two years' experience (in 1996–97), they still have the trainers present, but the trainers have left by the time they are observed with three or four years' experience (in 1998–99). Thus, the higher impacts of MES on student performance in schools with limited implementing experience may reflect the fact that MES trainers were present in those schools but not in the schools we observe with more implementing experience.

The timing of the MES trainers also coincides with our categories of student exposure. Considering only tests taken once schools have three or four years of experience with MES,[37] students with one or two years' exposure to MES entered an MES school in 1997–98 or 1998–99 and therefore attended the MES schools only after MES trainers had left. Furthermore, students with three years' exposure to MES entered in 1996–97 and overlapped with the MES trainers for one year and students with four years' exposure entered in 1995–96 and overlapped with the MES trainers for two years. Our student-exposure results suggest, therefore, that the impact of MES is not statistically significant without trainers, and that exposure to the trainers for two years has a larger impact than exposure to the trainers for one year. In short, the impact of MES appears to increase with student exposure, but only if trainers are involved.

CONCLUSIONS AND POLICY IMPLICATIONS

States around the country are now implementing school report cards and other accountability systems that are based on student test scores (Goertz & Duffy, 2001). This growing emphasis on student performance in education policy implies that test score data are becoming more widely available. This type of data provides an opportunity for scholars to use quasi-experimental methods to evaluate whole-school reform programs and other educational innovations. Because evaluations of this kind provide a valuable complement to studies based on random assignment, we hope that many scholars will take advantage of this opportunity.

The quasi-experimental evaluation in this paper reveals that the extensive efforts to implement whole-school reform in New York City have met with mixed success. To begin, we find no evidence that schools in New York City could increase their elementary reading test scores simply by adopting the School Development Program or Success for All. The results for SFA may surprise some readers because that model focuses on reading and because some previous studies have found evidence of positive impacts from SFA (Herman et al., 1999). We also find preliminary evidence, however, that the cumulative impact of SDP and SFA on student reading performance in grades 1 through 3 depends on the extent to which the SFA prescriptions are implemented and that these might have been positive, but not large, if the programs had been fully implemented.

In contrast, we find evidence that More Effective Schools had a positive impact on cumulative student performance in grades 1 through 3 and we find some evidence that it may also increase the value added to student performance in grade 4. Two considerations, however, cast some doubt on this evidence. First, the positive

[36] This information on MES implementation comes from our interviews with school personnel and the program developers. See Bifulco et al. (2002).

[37] Restricting comparisons to these schools ensures that the student-exposure effects do not reflect differences in schools' experience implementing school reform.

impacts estimates are not consistently distinguishable from zero in school level analyses. Second, we find some evidence that the positive impact of MES does not persist in the years when program trainers are no longer present.[38] This result suggests that schools may have difficulty maintaining the positive impact of MES on their own. Because the trainers require spending beyond the standard payments for school teachers and administrators, this result also suggests that whole-school reform may not be able to boost student performance unless it is accompanied by an increase in resources for more or better-trained personnel.

Overall, these results highlight the challenges facing poor, inner-city schools. We find evidence that whole-school reform may have a role to play in boosting student reading performance in these schools. Nevertheless, this potential contribution is undermined by key characteristics of these schools including: lack of resources; limited management and teaching skill, which lead to poor program implementation and the need for outside "trainers"; a concentration of students with limited English proficiency; and high student mobility. Further experiments with, and evaluations of, whole-school reform models are clearly warranted, but nobody should expect this approach to be a panacea for poor, inner-city schools.

We are grateful to the Smith-Richardson Foundation, which funded this research, and to Carolyn Bourdeaux, who played an important role in helping to collect and analyze the information on program implementation. We would also like to thank the New York City Board of Education Research Review Committee, which granted us access to our data; Jan Rosenbloom, who prepared the data files for us; and Dan Black and two anonymous referees, who gave us many helpful comments.

ROBERT BIFULCO is Assistant Professor of Public Policy at the University of Connecticut.

WILLIAM DUNCOMBE is Professor of Public Administration and Senior Research Associate at the Center for Policy Research at Syracuse University.

JOHN YINGER is Trustee Professor of Public Administration and Economics and Associate Director of the Center for Policy Research at Syracuse University.

REFERENCES

Barnett, W.S. (1996). Economics of school reform: Three promising models. In H.F. Ladd (Ed.), Holding schools accountable: Performance-based reform in education. Washington, DC: The Brookings Institution.

Bifulco, R. (2002). Addressing selection bias in quasi-experimental evaluations of whole-school reform: A comparison of methods. Evaluation Review, 26(5), 544–571.

Bifulco, R., Bordeaux, C., Duncombe, W., & Yinger, J. (2002). Do whole-school reform programs boost student performance? The case of New York City. (Final Report Submitted to the Smith-Richardson Foundation). Syracuse, NY: Syracuse University, Center for Policy Research.

Bloom, H.S., Ham, S., Melton, L., & O'Brien, L. (2001). Evaluating the accelerated schools

[38] As a referee pointed out to us, we cannot rule out the possibility that the disappearance of the MES effect is caused by some unobserved variable that changes in the same year the trainers leave the MES schools.

approach: A look at early implementation and impacts on student achievement in eight elementary schools. New York: Manpower Demonstration Research Corporation.

Bound, J., Jaeger, D.A., & Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association, 90(430), 443–450.

Cobb-Clark, D.A., & Crossley, T. (2003). Econometrics for evaluations: An introduction to recent developments. The Economic Record, 79 (December), 491–511.

Comer, J.P., Haynes, N.M., & Joyner, E.T. (1996). The school development program. In J.P. Comer, N.M. Haynes, E.T. Joyner, and M. Ben-Avie (Eds.), Rallying the whole village: The Comer process for reforming education. New York: Teachers College Press.

Cook, T.D., Habib, F., Phillips, M., Settersten, R., Shagle, S.C., & Degirmencioglu, S.M. (1999). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. American Education Research Journal, 36(3), 543–597.

Cook, T.D., Hunt, H.D., & Murphy, R.F. (1998). Comer's School Development Program in Chicago: A theory-based evaluation. WP-98-24. Evanston, IL: Institute for Policy Research, Northwestern University.

Emmons, C. (1999). School development program implementation report for New York community school district number 13. New Haven, CT: Child Study Center School Development Program, Yale University.

Ferguson, R., & Ladd, H.F. (1996). How and why money matters: A production function analysis of Alabama schools. In H.F. Ladd (Ed.), Holding schools accountable: Performance-based reform in education (pp. 265–298). Washington, DC: The Brookings Institution.

Goertz, M.E., & Duffy, M.C. (with K.C. Le Floch). (2001). Assessment and accountability systems in the 50 states: 1999–2000. (Consortium for Policy Research in Education CPRE Research Report Series RR 046, March 2001). Philadelphia, PA: University of Pennsylvania.

Goertz, M.E., & Edwards, M. (1999). In search of excellence for all: The courts and New Jersey school finance reform. Journal of Education Finance 25(1), 5–31.

Greene, W.H. (2003). Econometric analysis (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Heckman, J.J. (1979). Sample selection bias as a specification error. Econometrica 47(1), 153–161.

Heckman, J.J., LaLonde, R.J., & Smith, J.A. (1999). The economics and econometrics of active labor market programs. In A. Ashenfelter and D. Card (Eds.), Handbook of labor economics, (Volume 3A, pp. 1865–2097). Amsterdam: Elsevier.

Herman, R., Aladjam, D., McMahon, P., Masem, E., Mulligan, I., Smith, O., O'Malley, A., Quinones, S., Reeve, A., & Woodruff, D. (1999). An educator's guide to school wide reform. Arlington, VA: Educational Research Service.

Ladd, H.F., & Hansen, J.S. (1999). Making money matter: Financing America's school. Washington DC: National Academy Press.

Millsap, M.A., Chase, A., Obiedallah, D., & Perez-Smith, A. (2001), Evaluation of the Comer School Development Program in Detroit, 1994–1999: Methods and results. (Paper presented at the annual meetings of the Association for Public Policy Analysis and Management). Washington, DC.

Northwest Regional Educational Laboratory (NWREL). (1998). Catalog of school reform models (1st ed.). Washington, DC: U.S. Department of Education.

Raudenbush, S.W., & Bryk, A.S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage Publications.

RMC Research Corporation (RMC). (1976). Interpreting NCEs. (Technical Paper No. 2). Mountain View, CA: RMCC Research Corporation Mountain View.

Rouse, C.E. (1998). Private school vouchers and student achievement of the Milwaukee Parental Choice Program. Quarterly Journal of Economics 113 (2), 553–602.

Schwartz, A.E., Stiefel, L.S., & Kim, D.Y. (2004). The impact of school reform on student performance: Evidence from the New York Network for School Renewal Project. Journal of Human Resources, 39 (2), 500–522.

Slavin, R.E., Madden, N.A., Dolan, L.J., & Wasik, B.A. (1996). Every child, every school: Success for all. Newbury Park, CA: Corwin.

Stata Corporation. (2003). Stata user's guide, release 8. College Station, TX: Stata Press.

Viadero, D. (2001). Memphis scraps redesign models in all its schools. Education Week, 20(42), 1–19.

Wooldridge, J.M. (2002). Econometric analysis of cross section and panel data. Cambridge, MA: MIT Press.

Table A-1. Value-added Production Functions Estimates for Fifth Graders in 1997 and 1999, with Heckman Selection Correction.

	SDP		MES		SFA	
	OLS	IV	OLS	IV	OLS	IV
N	9,189	9,189	7,857	7,857	7,150	7,150
Uncensored observations	8,261	8,261	6,899	6,899	6,393	6,393
I. The Production function						
Adopted whole-school reform	–0.119	–0.302	–0.008	0.869	–2.315**	–3.269
[Standard error]	[0.755]	[1.519]	[0.647]	[2.610]	[0.641]	[3.327]
Individual characteristics						
Year = 1999	–0.167	–0.98	–2.017	–1.526	–1.323	–1.279
Lagged test-score	0.643**	0.626**	0.627**	0.619**	0.628**	0.621**
Lagged test–score if > 50	0.036**	0.042**	0.035**	0.043**	0.038**	0.043**
Lagged test-score*year = 1999	–0.010	0.012	0.035	0.026	0.030	0.024
Lagged test-score if > 50*year = 1999	–0.022	–0.034*	–0.041**	–0.040**	–0.038**	–0.039**
Female	0.694**	0.454	0.828**	0.476	0.849**	0.580
Asian (reference category is white)	4.872**	7.767**	4.590**	6.930**	3.767**	5.916**
Hispanic (reference category is white)	–1.005	–0.698	–1.125	–0.700	–0.702	–0.536
Black (reference category is white)	–2.086**	–1.935**	–2.326**	–2.006**	–1.908**	–1.754*
Free lunch eligible	–1.128**	–0.574	–1.263**	–0.887	–1.660**	–0.979
Eligible for ESL services[a]	–1.232	5.789	–0.704	4.438	–3.031**	5.193
Lamba (inverse Mills ratio)	–1.357	–20.833	–3.502**	–16.101	2.648	–18.866
School characteristics						
Log of enrollment*10	0.112	0.160	0.306**	0.218*	0.142	0.174
% free lunch	0.027	0.051	0.040	0.061	0.066	0.074
% limited English proficient	0.020	0.012	–0.015	–0.005	0.023	0.019
% Hispanic	–0.023	–0.019	–0.007	–0.011	–0.015	–0.016
% overage for grade	–0.059	–0.025	–0.001	–0.005	–0.030	–0.012
% teachers < 2 yrs experience	–0.064	–0.024	–0.008	–0.003	–0.002	–0.001
% teachers w/certification	0.009	0.041	0.038	0.055	0.077	0.072
Average class size	–0.086	–0.089	–0.266*	–0.135	–0.066	–0.083
SURR[b]	–0.847	–0.700	–0.999*	–0.765	–0.581	–0.468

(continued)

Table A-1. (continued)

	SDP	MES	SFA
II. The Selection equation[c]			
Female	0.060*	0.085**	0.048
Asian	–0.393**	–0.381**	–0.166**
Free lunch eligible	–0.200**	–0.171*	–0.138**
Eligible for ESL services[a]	–0.973**	–1.016**	–0.998**
Home language other than English	–0.073*	–0.119**	–0.071*

* = significant at the 0.10 level. ** = significant at the 0.05 level. All inferences based on robust standard errors.
[a] = 1 if student was eligible for English as Second Language (ESL) services during the previous school year, 0 otherwise.
[b] = 1 if school under registration review during the outcome year, 0 otherwise.
[c] This panel gives results for the first-stage regressions in the Heckman-selection procedure for each reform program.