ISSN: 1525-3066

Center for Policy Research
Working Paper No. 55

Education Finance and Accountability Program
(EFAP)

DOES WHOLE-SCHOOL REFORM BOOST STUDENT PERFORMANCE? THE CASE OF NEW YORK CITY

Robert Bifulco, William Duncombe, John Yinger*

Center for Policy Research
Maxwell School of Citizenship and Public Affairs
Syracuse University
426 Eggers Hall
Syracuse, New York 13244-1020
(315) 443-3114 | Fax (315) 443-1081
e-mail: ctrpol@syr.edu

August 2003

$5.00

CENTER FOR POLICY RESEARCH – Summer 2003

Timothy Smeeding, Director
Professor of Economics & Public Administration

Associate Directors

Margaret Austin
Associate Director,
Budget and Administration

Douglas Holtz-Eakin
Professor of Economics
Associate Director, Center for Policy Research

Douglas Wolf
Professor of Public Administration
Associate Director, Aging Studies Program

John Yinger
Professor of Economics and Public Administration
Associate Director, Metropolitan Studies Program

SENIOR RESEARCH ASSOCIATES

Dan Black ..... Economics
Stacy Dickert-Conlin ..... Economics
William Duncombe ..... Public Administration
Gary Engelhardt ..... Economics
Deborah Freund ..... Public Administration
Madonna Harrington Meyer ..... Sociology
Christine Himes ..... Sociology
William C. Horrace ..... Economics
Bernard Jump ..... Public Administration
Duke Kao ..... Economics
Eric Kingson ..... Social Work
Thomas Kniesner ..... Economics
Jeff Kubik ..... Economics
Andrew London ..... Sociology
Len Lopoo ..... Public Administration
Jerry Miner ..... Economics
John Moran ..... Economics
Jan Ondrich ..... Economics
John Palmer ..... Public Administration
Lori Ploutz-Snyder .. Health and Physical Education
Jeff Racine ..... Economics
Grant Reeher ..... Political Science
Stuart Rosenthal ..... Economics
Ross Rubinstein ..... Public Administration
Michael Wasylenko ..... Economics
Janet Wilmoth ..... Sociology

GRADUATE ASSOCIATES

Anna Amirkhanyan ..... Public Administration
Beth Ashby ..... Economics
Dana Balter ..... Public Administration
Gabby Chapman ..... Economics
Yong Chen ..... Economics
Christopher Cunningham ..... Economics
Ana Dammert ..... Economics
Tae Ho Eom ..... Public Administration
Val Episcopo ..... Sociology
Mike Eriksen ..... Economics
Kate Farrar ..... Public Administration
Garey Fuqua ..... Public Administration
Jose Galdo ..... Economics
Andrzej Grodner ..... Economics
Glenda Gross ..... Sociology
Anil Kumar ..... Economics
Kristina Lambright ..... Public Administration
Xiaoli Liang ..... Economics
Liqun Liu ..... Economics
Joseph Marchand ..... Economics
Cristian Meghea ..... Economics
Desmond Nation ..... Public Administration
Emily Pas ..... Economics
Seth Richards ..... Public Administration
Cynthia Searcy ..... Public Administration
Claudia Smith ..... Economics
Sara Smits ..... Sociology
Adiano Udani ..... Public Administration
Wen Wang ..... Public Administration
Bo Zhao ..... Economics
Na Zhao ..... Sociology

STAFF

Kelly Bogart ..... Administrative Secretary
Martha Bonney ..... Publications/Events Coordinator
Karen Cimilluca ..... Librarian/Office Coordinator
Kim Desmond ..... Administrative Secretary
Kati Foley ..... Administrative Assistant, LIS
Kitty Nasto ..... Administrative Secretary
Candi Patterson ..... Computer Consultant
Mary Santy ..... Administrative Secretary
Mindy Tanner ..... Admin. Secretary/Receptionist

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—do not significantly increase reading scores but might have if they had been fully implemented. The More Effective Schools program does boost reading scores, particularly for the poorest students, but only when program "trainers" remain in the school and the students are native English speakers.

Key words: Elementary education, whole-school reform, quasi-experimental evaluation.

Introduction

Student performance in school districts with concentrated poverty, particularly large city districts, is far below student performance in other districts. In Baltimore, New Orleans, and Philadelphia, for example, the share of students scoring above a selected level on eighth-grade reading and math tests falls more than 50 percent below the state average (Casserly, 2002). 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 decision making, resource allocation, classroom organization, curriculum and instruction, parental involvement, and student support. Traditional school-level interventions usually focus on one of these issues.[1]

Efforts to implement whole-school reform have been accelerating, particularly in urban schools that serve disadvantaged and minority students. The Comprehensive School Reform Demonstration program (CSRD), enacted by Congress in 1997 and re-authorized 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 and 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 and Hansen 1999: 153).

More recent contributions to the literature include evaluations of SDP in Prince George County, Maryland (Cook, *et al.* 1999), Chicago (Cook, *et al.* 1999), 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, and Kim forthcoming), 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 presents results from a quasi-experimental study of whole-school reform efforts undertaken in New York City during a three-year period 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.

This study cannot make use of random assignment, a strategy with well-known advantages for deriving estimates of program impacts. For several reasons, however, researchers cannot rely solely on randomized assignment to evaluate whole-school reform models. Because these models involve the whole school, researchers cannot randomly assign individual students or teachers within a school. Thus, an experiment must identify a set of schools interested in whole-school reform and then randomly deny some of them the opportunity to implement a whole-school reform plan. Because of the difficulty of recruiting schools willing to agree to these conditions, experimental studies are unlikely to provide precise estimates of program impacts.[3] More importantly, because studies based on random assignment typically are small in scale with active participation by program developers, the treatment schools in these studies usually receive more attention than would the average school in any large-scale effort to implement whole-school reform. Thus, experimental studies are unlikely to reveal whether policies that encourage or mandate whole-school reform in a large number of schools can be expected to foster consistent improvement.

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 challenge in using quasi-experimental data to estimate program impacts arises because treatment schools are likely to be self selected. If unobserved factors that influence either a school's decision to adopt a whole-school reform or students' decisions to attend a school that has adopted whole-school

reform also influence school and student performance, then cross-sectional comparisons of adopting schools with non-adopting schools might provide biased estimates of model impacts. Our methodology, which is discussed in a later section, addresses this issue 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.

The paper has four main sections. The first section describes our sample and our data. The second section explains our estimation strategy. Using a standard education production function, we derive the possible sources of bias in estimating program impacts, explain our strategy for eliminating these biases, and show how our data can be used to implement our approach. The third section contains our results. We describe our main findings, explore a variety of alternative estimation strategies, and address several issues that arise in our earlier discussion of methodology. The final section presents conclusions and policy implications.

Data

The Study 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).[4] 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).[5] 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.[6]

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 1994-1995. 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.

Stratified random sampling was used to select additional schools to serve as a 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.[7] Next, we created three sampling frames corresponding to the three years in which whole-school reform models were adopted. Each frame was split into quartiles based on student performance, and an equal number of schools was randomly selected from each 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.[8] Overall, 28 schools were selected from the 1994-1995 sampling frame, 12 from the 1995-1996 frame, and 12 from the 1996-1997 frame. Some schools were selected from more than one sampling frame. In addition, we dropped two of the selected schools because they said they had adopted a whole-school reform model in either 1997-1998 or 1998-1999. The final sample contains 40 comparison schools.

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.[9] 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).[10] The NCE measure can be interpreted as an equal-interval scale; with a normally distributed 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. Because NCEs form an equal-interval scale, they can meaningfully be aggregated and averaged (RMC Research Corporation (RMC), 1976).

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.[11] 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).

Table 3 compares 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 proceeding the reference year used for the earliest sampling frame from which they were selected. This table shows 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 last two rows of Table 3 show the percentage of third grade students who scored above the statewide reference point (SRP) on the New York State Pupil Evaluation Program (PEP) tests in reading and math.[12] These pre-adoption performance measures are similar across all groups of schools, except that SDP schools show a higher average percentage of students above the SRP in third-grade reading.

Estimation Strategy

While our treatment and control group schools are a fairly close match on observable characteristics, important unobservable differences could exist between them. In this section, we discuss in detail the strategy we use to control for potential sample selection bias associated with such unobservable differences.

Educational Production Functions

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 λ. 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.

This production function cannot be estimated in this form because neither the student and school fixed effects nor, in most cases, the explanatory variables in previous years can be observed. Moreover, estimating it with only observable information, that is, the contemporaneous X's and W, may result in biased estimates because the omitted variables may be correlated with the observable variables.

The potential bias we are most interested in, of course, involves β. The estimate of β may be biased if unobserved characteristics of a school, which are included in δ and in lagged values of the school-level X's, are correlated with the decision to adopt whole-school reform. This type of bias is sometimes called self-selection bias because it arises when schools with certain unobserved traits are more likely to select reform. 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.

Strategies to Eliminate Bias

Several methods have been developed to eliminate these potential biases. 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. 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:

$$Y_{ijT} = \alpha X_{ijT} + \beta W_{ijT} + \lambda Y_{ijT-1} + (\gamma_t - \lambda\gamma_{t-1}) + (\varepsilon_{ijT} - \lambda\varepsilon_{ijT-1}). \qquad (2)$$

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.

A technical problem arises with this equation because the lagged dependent variable is correlated with the component of the error term that has the same time subscript. Some scholars call this the problem of "measurement error" in test scores, because the lagged test score may not accurately measure the impact of previous *X*s. In any case, this correlation can lead to biased results. A standard way to eliminate this bias is to use a two-year lagged test score as an instrument for the one-year lagged test score.[14] Using the subsample for which this approach is feasible, we check for the presence of this type of bias in our data.

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

$$Y_{ijT} - Y_{ijT-1} = \alpha(X_{ijT} - X_{ijT-1}) + \beta(W_{ijT} - W_{ijT-1}) + (\gamma_t - \gamma_{t-1}) + (\varepsilon_{ijT} - \varepsilon_{ijT-1}), \qquad (3)$$

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 answering our second research question.[16]

Another approach, and the one we prefer, is to estimate equation (2) using an instrumental-variables (IV) technique that accounts for the potential impact of unobserved school characteristics, δ, on the decision to adopt whole-school reform, W. 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. [17]

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 (i.e., 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 or in any data set we know of.[18]

Fortunately, however, we can answer our first research question using our IV technique without the lagged dependent variable, that is, without 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 X's 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 standard errors of the coefficients, but this variable is not required to eliminate self-selection bias in β.[19]

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 (Woolridge, 2003). 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 all the unobservables in equation (1).

The link between our data and our methods 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 1994-1995 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 1996-1997 cohort, on the other hand, students in schools that adopted whole-school reform in 1994-1995 experienced whole-school reform starting in the first

grade, so they have spent their entire early elementary years in a whole-school reform school by the time we observe their third grade scores in 1996-1997.

The last two columns of Table 4 indicate the number of treatment schools (for each reform model) and the number of observations (i.e. 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 school 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 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 during 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] If student mobility rates are different in treatment and comparison schools and students who change schools show different rates of performance growth, controlling for other differences, then dropping movers from the sample could introduce an additional source of bias.

In sum, excluding students with missing test scores or students who have moved to schools outside the study sample may bias estimates of model impacts.[22] To address this potential bias, we employ a Heckman two-step selection-correction procedure (Heckman 1979). To further check the sensitivity of impact estimates to the exclusion of movers, we conduct alternative analyses in which movers are included.

Definition of Treatment

Schools vary in the success with which they implement a particular model of whole-school reform. Moreover, the principles and practices associated with many models have diffused beyond the schools that have made an explicit decision to adopt a whole-school reform model. Under these circumstances, it is not obvious how to define and measure the intervention that whole-school reform represents.

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). As in that case, two approaches are possible. The first is to focus on the decision to adopt a reform and the second is to focus on the degree to which the specific reform prescriptions are adopted.

For several reasons, we focus on the impact of the decision to adopt a reform. First, the decision to adopt is subject to direct policy control of a school or school district, whereas the extent to which policy prescriptions are implemented also reflects the efforts and abilities of teachers, other school staff, and even parents. Second, schools that do a good job implementing a model may not be representative of either the schools adopting that model or the population of schools targeted for future interventions. Thus, focusing on the impact of well-implemented model components limits the ability to generalize any findings. Third, the extent to which model prescriptions are followed in a school can be difficult to measure. Finally, factors that influence the quality of implementation might be more closely related to student performance than are the factors influencing the decision to adopt. If so, self-selection biases might be more difficult to eliminate when estimating the impacts of model implementation than when estimating the effects of the adoption decision.

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 W depends on the quality of implementation, based on information provided by the program developers.

The impact of whole-school reform also might depend upon a school's experience with a particular whole-school reform model. Because the extensive organizational changes required by whole-school reform may take a number of years to achieve, for example, a finding that student performance has not improved after one or two years does not necessarily imply that the program has been or will be ineffective. On the other hand, the enthusiasm surrounding initial efforts at

school reform might engender short-term gains that are not maintained in later years. To account for these possibilities, we determine whether program impacts vary with the length of time a school has been implementing a particular reform model.

Finally, student mobility implies that not all students in a treatment school have been exposed to whole-school reform for the same number of years. Our basic results indicate the impact of whole-school reform on all the students in treatment schools at a given point in time. In fact, however, a reform program's impact may increase with the length of time a student has been exposed to it. To account for this possibility, we also determine whether program impacts are influenced by 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] To illustrate the impact of our decision to use IV estimators, these tables (and the following ones) present both OLS and IV results. Our discussion will focus exclusively on the IV results.

Table 5 presents results for our first question, which concerns the cumulative impact of whole-school reform in grades one through three. Our preferred results, which appear in the fourth column, are based on IV estimation with a Heckman selection correction (Heckman, 1979).[24] As explained earlier, this procedure controls for the impact that a whole-school reform program might have on test scores through its influence on whether or not a score is observed. These results indicate that SDP and SFA have small, insignificant impacts, whereas MES has a large, positive impact on reading performance. The NCE scale is designed to have a standard deviation (SD) of 21.06. Consequently, the point estimate in the fourth column of Table 5

implies that three years of exposure to MES raises the average student's test score by 61.4 percent of an SD, which is a large impact.

The results for question two, concerning value-added impacts, are presented in Table 6. Our preferred results, which are presented in the second panel, also indicate a positive contribution by MES. In particular, MES raises the fourth-grade value-added by 22.7 percent of an SD. However, MES does not have a significant impact on fifth-grade value-added. The value-added impacts for SDP and SFA are small and insignificant for all grades.

Tables 5 and 6 also present several other sets of results corresponding to alternative ways of dealing with the methodological concerns discussed earlier in the paper. The first two columns in Table 5 and the first panel in Table 6 present results without a selection correction for students who did not leave sample schools. The last two columns in Table 5 and the third panel of Table 6 present results without a selection correction for all students, including those who moved out of sample schools. These results are very similar to our preferred estimates; indeed, they exhibit the same pattern of statistical significance and have virtually the same magnitudes. Our results clearly are not driven by the way we handle missing test scores or student mobility.

Another issue is that measurement error in a lagged test score 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 instrumental variables procedure, with test score lagged two years as the instrument, to estimate our model for the subsample with the required data. The results are presented in the fourth panel of Table 6. Although the fifth-grade result for SDP changes signs with this procedure, it is still small and insignificant, and the fifth-grade results for MES and SFA are very similar to those in the first panel of the table. Thus, our results do not appear to be affected by this measurement error.

Finally, we test the validity of our IV procedure by comparing its results with the results of more precise procedures that combine the value-added and differencing approaches. As

explained earlier, this comparison is only possible for a selected subsample of our data. For this subsample, Table 7 shows that our procedure (VA-IV) provides estimates of fifth-grade, value-added impacts similar to those obtained from more precise approaches (D-VA and D-VA-IV).[25] Thus, our IV procedure appears to provide sufficient protection against selection bias.

Do Impacts Vary with Student Characteristics?

The impact of whole-school reform may not be the same for all types of students. Exploration of this type of variation in model impacts is crucial to identify the types of schools and students for which a particular model is best suited. We investigate whether model impacts depend on ethnicity, poverty, or English proficiency. Our results are presented in Table 8.

For SDP, we can reject the hypothesis that the cumulative impact of whole-school reform through grade three is the same for black and Hispanic students. We can also reject the hypothesis that these two groups have the same value-added impact in fourth grade. In both cases, however, the underlying estimates of program impact are not statistically significant.

Some stronger results appear for MES. As shown in the second panel of Table 8, MES has a significant positive impact for poor students (as measured by eligibility for a free lunch) and an insignificant negative impact for non-poor students. In addition, MES has a large, positive, significant impact on the performance of native speakers but a much smaller, insignificant impact on students who speak English as a second language. The difference between these results is statistically significant. Overall, MES is successful in overcoming the disadvantages associated with poverty but not those associated with a need to learn English.

The third panel of Table 8 provides results for SFA. None of the results in this panel indicate significant differences in program impacts for different types of students.

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 a

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 which implementation was less successful. 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.[26]

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 of the sample mean for schools in a given year.[27] 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 9. We find that cumulative program impacts do 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 a SD in 1997 and one-fifth of a SD in 1999.[28] A comparable improvement in implementation for SDP would have a smaller impact, less than one-seventh of an SD.

Because our implementation measures are exploratory and because these regressions are not estimated with an IV procedure, these results are no more than suggestive. With this caution in mind, they indicate that implementation matters and that these two programs are unlikely to have substantial benefits unless implementation is virtually perfect. This finding reinforces the value of quasi-experimental studies, which are more likely than studies based on random assignment 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 New York City, where implementation was difficult to control.

Do Model Impacts Vary with Length of Time Exposed?

As explained earlier, 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. Our data set is ideal for exploring these possibilities. In many cases, we observe the performance on the same test in the same treatment schools for students in different cohorts. Moreover, because of student mobility, we always observe (except in the first year a program is implemented) students in treatment schools with different years of exposure to a reform program.

In the case of school experience, we can determine whether the two years separating cohorts, which corresponds to two more years of experience with whole-school reform in treatment schools, boosts program impacts. In schools that implemented whole-school reform in 1995-96, for example, we observe fourth-grade test scores in 1995-96, when these schools had one year of experience with the program, and in 1997-98, when they had three years of experience. Our approach is to compare the program impacts in these two years.

Our results are presented in Table 10. The first panel presents results for cumulative impacts. The first two columns apply to schools with one or two years experience and the second two columns apply to schools with three or four years experience.[29] The results are striking. For every reform model, the cumulative impact is larger for schools with fewer than three years of experience than for schools with three or more years of experience. Because these cumulative impacts are based on third-grade test scores, however, students in schools implementing for fewer than three years have not been exposed to the reform program for their entire early elementary education. Consequently, the results in the first two columns differ from those in the next two columns both because the schools have less experience with reform and because the students have less exposure to it. Because these results mix school experience and student exposure, they cannot be interpreted without examining whether performance gains are influenced by the length of time a student has been exposed to a reform program.

Our procedure for studying student exposure is to interact the treatment variable with the number of years the student has been exposed to the reform program. This procedure yields a separate estimate of program impact for each length of exposure. To ensure that our results do not reflect differences in each school's experience implementing whole-school reform, we base this analysis only on schools with three or four years experience with a reform program.

The results are presented in Table 11. As shown in the first panel of this table, we find some significant differences in the cumulative impacts of whole-school reform through third grade based on student exposure. For SDP, we obtain the curious result that the program has significant, positive impact for students exposed for a single year. Indeed, this impact is larger than the impact for students exposed from two to four years. However, this result is only significant at the 10 percent level.

For MES, we find that more exposure appears to indicate a larger cumulative impact. The impact is about the same for students exposed for one or two years, but is larger for student exposed for three years, and larger still for students exposed for four years, which implies that they entered the program in kindergarten. The results for three and four years of exposure are statistically significant at the 5 percent level (and significantly different from each other). These results suggest that more years of exposure leads to a greater program impact.

Considering the results of Tables 10 and 11 together, leads to a more complete explanation. In the case of SDP, the only significant impacts arise with limited student exposure (Table 11) and the combination of limited student exposure and limited school experience (Table 10). Taken literally, these results indicate that the gain from limiting student exposure to SDP is magnified by limiting the school's experience with SDP. It is as if the program generates some initial enthusiasm, on the part of both students and schools, that has a significant impact on student performance, but this enthusiasm (and its performance impact) fades as contact with the program continues.

In contrast, the MES results in Tables 10 and 11 seem to work in opposite directions. Table 11 indicates that program impacts increase with student exposure, whereas Table 10 indicates that program impacts are greatest when school experience is lowest, despite the fact that limited school experience coincides with limited student exposure. A plausible interpretation of this result is suggested by a key feature of MES implementation in the New York City schools. Specifically, the people who developed MES 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 any of the MES schools in 1997-98 or 1998-99.[30] This timing coincides with the categories in Table 10. In particular, when schools are observed with fewer than two years experience (in 1996-1997), they still have the trainers present, but the trainers have left by the time they are observed with three or four years experience (in 1998-1999). Thus, it is the presence of the MES trainers, not a lack of school experience, which probably generates the positive impact of MES on student performance.

This timing also coincides with the categories in Table 11. Recall that the results in this table are based only on tests taken once schools have three or four years of experience with MES. Because MES was not adopted in any schools in 1994-95, this implies that the MES results in this table are based exclusively on third-grade tests taken in 1998-99. Students with one or two years exposure to MES in this subsample 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. The results in Table 11 indicate, therefore, that the impact of MES is not statistically significant (at the 5 percent level) 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 increases with student exposure, but only if trainers are involved.

Table 10 also explores the impact of school experience on the value-added impact of the three programs. As shown in the second and third panels of this table, no clear patterns emerge. The point estimates indicate that SFA has a negative impact in both fourth and fifth grades when it is first implemented, but these results are not statistically significant. Indeed, the only significant impact is for MES in schools with fewer than two years experience. The estimated MES impact is larger for schools with three or four years experience, but it is not significant. These value-added impacts for MES may be harder to sort out than the cumulative impacts because the categories no longer have a clear link to the MES trainer issue; in fact, all the schools in the value-added regressions have trainers for at least one year.[31]

The second panel of Table 11 hints that the value-added impact of whole-school reform actually declines with student exposure to the program. Indeed, the point estimates are consistent with this view for all three reform models. The results in this panel are not statistically significant, however. The regressions on which these results are based pool fourth and fifth grade test scores to ensure that there are enough students in each experience category to make the estimation feasible. Even with this pooling, however, we are unable to identify any significant impact of student exposure on programs' value-added impact.

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 and 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, along with Schwartz, Stiefel, and Kim (forthcoming), 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 Success for All boosted elementary reading test scores in New York City. This result may surprise some readers because SFA focuses on reading and because some previous studies have found evidence of positive impacts from SFA (Herman, et al. 1999). Nevertheless, we find that SFA does not have a substantial or statistically significant impact on either cumulative student performance in grades one through three or on value-added to student performance in grades three, four, or five.

We also find that the cumulative impact of SFA on student performance in grades one through three depends on the quality with which the SFA prescriptions are implemented. It follows that SFA's cumulative impact might have been larger if the program had been better implemented. The average implementation rating of the SFA schools was quite high, however, so our estimates imply that the cumulative impact of SFA would still not have been very large except in the unlikely event that every school exactly followed all of SFA's prescriptions.

The results for the School Development Program also are discouraging. Although our point estimates indicate that SDP has a more positive impact on reading scores than does SFA, its impacts are also small and statistically insignificant. We also obtain similar implementation results for SDP as for SFA. Better implementation would boost SDP's cumulative impact over grades one through three, it appears, but our results imply that even with complete implementation, the cumulative impact of SDP would be small.

In contrast, we find that More Effective Schools has a large and statistically significant impact on cumulative student performance in grades one through three and on the value added to

student performance in grade four. Indeed, we find that MES raises reading test scores by 84.1 percent of a standard deviation over the first four years of school, which is a surprisingly large impact. Moreover, MES appears to be more effective for poor students than for non-poor students, a characteristic of great value in an inner-city setting. It also appears to be more effective for native English speakers, both black and Hispanic, than for students who speak English as a second language.

One key feature of MES is that the program developers send trainers to work with each adopting school for the first couple years of implementation. We find that these trainers play a key role; in fact, the large, positive impact of MES disappears when the MES trainers stop coming. This is, of course, a discouraging result. It suggests that schools 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 indicates that whole-school reform may not be able to boost student performance unless it is accompanied by a significant 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.

Endnotes

* The authors are, respectively, Assistant Professor of Political Science, University of Connecticut; Professor of Public Administration, The Maxwell School, Syracuse University; and Professor of Economics and Public Administration, The Maxwell School, Syracuse University. 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, and to Jan Rosenbloom, who prepared the data files for us.

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

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

3. The largest number of schools in an experimental study is 23, including 13 School Development Program (SDP) schools, in Cook et al. (1999). Cook, Hunt and Murphy (1998) examine 19 schools, including 10 SDP schools. In addition to small scale, studies based on random assignment, also face a threat from teacher movement in or out of schools that adopt a reform model.

4. For brief descriptions of SDP and SFA see Barnett (1996) or NWREL (1998). For more complete descriptions see Comer, Haynes and Joyner, 1996 and Slavin *et al.*, 1996. For a description of MES see the Association for Effective Schools, Inc. website at http://www.mes.org.

5. 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.

6. Several schools in the control group are also SURR schools. See Table 4.

7. These districts serve few poverty students in comparison with districts that have 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 endnote 5).

8. 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 (endnote 5), which helps ensure that they will yield a comparison

group with a distribution of pre-adoption performance similar to that in the treatment schools.

9. Different tests measure different dimensions of reading performance and may use different norming procedures and samples. In 1994-1995 the NYCBOE used *Degress of Reading Power*, a test of reading comprehension developed by Touchstone Applied Science Associates. In 1995-1996, NCYBOE switched to a reading test published by CTB/McGraw-Hill, and in 1998-1999 began to use the reading component of the *TerraNova CAT*.

10. 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.

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

12. The PEP test and associated SRP, which is a minimum competency standard, were used until 1998-1999 to identify students for remedial assistance. Our data set includes PEP scores, but only at the school level, so we cannot use them in our empirical analysis.

13. The combination of school-level and individual-level variables in any production function study, including ours, calls for the use of robust standard errors. As a result, all of our standard errors are calculated using the "cluster" option in STATA, which makes use of a generalization of the Huber/White/Sandwich estimator of variance.

14. Schwartz, Stiefel, and Kim (forthcoming) deal with this problem by including both a one- and a two-year lagged values of test scores in the equation. This approach does not eliminate the correlation that causes the problem.

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 $\delta j(1-\lambda)$. Subtracting equation (1) for T-1 from equation (1) results in an equation (3) with the following new terms on the right side: $ST\text{-}1(\lambda\text{-}1)+\lambda(\alpha \text{XijT-}1 + \beta WijT\text{-}1)$, where *St* is the sum in equation (1) for year *t*.

16. A later section describes how we use this subset. As discussed below, pre-implementation test scores are rarely, if ever, available for studies of program impacts in early elementary school.

17. Schwartz, Stiefel, and Kim make a similar argument (page 14 of manuscript, endnote 10): "Students or their parents are unlikely to have known about this reform before enrolling (or even after enrolling)."

18. 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.

19. As before, this IV technique does not rule out the possibility of bias from a correlation between μ and parental decisions about where to send their children to school. In this case it also does not rule out bias from a correlation between these decisions and lagged individual characteristics in X. For the reasons given earlier, however, it seems likely that both of these correlations and the resulting bias are close to zero.

20. The percentage of students missing the test scores needed for the specific analyses presented here is less than these figures, and varies by cohort and school year.

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 (forthcoming), who retain in their sample all students who remained in any New York City school.

22. Schwartz, Stiefel, and Kim (forthcoming) do not correct for this potential selectivity problem. Because they include movers in their regressions, the selectivity problem in their case arises solely because some test scores are missing.

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 with a Heckman selection correction. Results for any other regression in this paper are available from the authors upon request.

24. In particular, we estimate an equation to predict 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. Tables 5 and 6 indicate that the Heckman selection term (called lambda or the inverse Mills ratio) is not significant in the fifth-grade regressions. Moreover, it is significant in only one of the IV regressions. It is significant, however, in most of the OLS regressions.

25. The value-added, difference-in-difference approach does not account for the possibility that schools or students have unobserved time trends in their test scores and that these trends might be correlated with model adoption. These trends, which are not included in equation (1), can be accounted for with double differencing. See Bloom (1984). We do not have enough data to implement this approach, but the D-VA-IV model in Table 7 is a step in this direction.

26. These ratings come from unpublished surveys supplied to us by the program developers. More information on the implementation ratings for SDP and SFA is available from the authors upon request.

27. 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.

28. 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 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 9).

29. We also observe cumulative impacts and fifth grade impacts for 25 SDP schools with five years experience. These estimates, which apply to schools that implemented SDP in 1994-95, are all small and insignificant and are omitted to simplify the presentation. We also observe two SFA schools with five years experience, but that is not enough to obtain reasonable estimates.

30. This information on MES implementation comes from our interviews with school personnel.

31. Because many of these value-added results are based on a small number of schools, we also estimated school-experience regressions that pooled fourth and fifth grade test scores. This approach makes it possible to include all treatment schools but also implies that fourth-grade test score gain for 1994-95 cohort (when the school has limited experience with a program) is sometimes compared to the fifth-grade test score gain for the 1996-97 cohort in the same school when it has more experience. With this approach, the negative impact in SFA schools with limited experience is statistically significant, but only at the 10 percent level. No other results are statistically significant. These results are available from the authors upon request.

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

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

Table 3. Means and Standard Deviations for Schools in the Study Sample[a]

	Adopters			Comparison
	SDP	MES	SFA	Schools
Number of schools	28	10	9	40
Number of SURR schools[b]	15	9	5	15
Enrollment	753 [273]	1050** [348]	886 [242]	751 [300]
% asian	0.6 [0.9]	1.0 [1.0]	1.5 [1.4]	0.8 [1.4]
% black	67.4** [28.5]	32.7** [29.0]	60.2 [18.8]	51.6 [30.1]
% hispanic	30.0** [27.1]	64.8** [29.2]	37.0 [17.4]	45.5 [28.6]
% white	1.8 [2.9]	1.4 [3.1]	0.9 [0.9]	1.9 [4.0]
% limited English proficient	13.6 [13.1]	32.8** [23.2]	19.1 [13.3]	18.5 [14.5]
% eligible for free lunch	87.8** [8.4]	93.7 [6.6]	94.1 [5.8]	91.8 [7.6]
Average class-size	27.4 [2.5]	28.4 [3.6]	27.4 [3.1]	27.6 [2.6]
% teachers <2 years experience	12.1 [7.1]	12.1 [4.5]	7.9 [5.7]	11.1 [7.8]
% teachers certified in field of assignment	79.5 [9.9]	76.3 [9.9]	90.1* [6.5]	80.6 [12.6]
% above SRP on grade 3 PEP reading	51.9* [16.2]	45.7 [14.7]	46.9 [11.8]	46.5 [13.8]
% above SRP on grade 3 PEP math	78 [11.6]	83.4 [7.6]	80.7 [5.7]	80.3 [7.8]

a. Reported averages [and standard deviations] are for the last year prior to program adoption. In the case of comparison schools, figures are from the year preceding the reference year used to define the earliest sampling frame from which the school was selected; * Indicates significantly different than the comparison group mean at the 0.10 significance level; ** Indicates significantly different than the comparison group mean at the 0.05 significance level.
b. Counts all schools that have been designated as a registration review school at any time.

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)	3,353 [5,685]
	1998-99	1994-95 1995-96 1996-97	10 (MES)	855 [5,685]
			9 (SFA)	885 [5,685]
Question 2: Value-Added Impact				
Grade 3	1994-95	1994-95	25 (SDP) 0 (MES) 2 (SFA)	1,827 [2,771] 0 0
Grade 4	1994-95	1994-95 1995-96	28 (SDP)	3,483 [5,993]
			10 (MES)	1,511 [5,993]
	1996-97	1994-95 1995-96 1996-97	9 (SFA)	1,208 [5,993]
Grade 5	1994-95	1995-95 1995-96 1996-97	28 (SDP)	3,156 [5,105]
			10 (MES)	1,794 [5,105]
	1996-97	1994-95 1995-96 1996-97	9 (SFA)	1,288 [5,105]

TABLE 5. Estimates of the Cumulative Impact of Whole-School Reform Through Grade 3, Using Alternative Samples and Specifications[a]

	Basic Estimates		With Heckman Selection Correction[b]		Including Movers[c]	
	OLS	IV	OLS	IV	OLS	IV
SDP	1.144	2.333	**1.141**	2.340	0.980	2.033
	[1.437]	[2.235]	**[1.427]**	[2.236]	[1.339]	[2.138]
MES	2.878**	12.939**	**2.819****	12.933**	2.660**	11.399**
	[1.437]	[3.601]	**[1.437]**	[3.602]	[1.291]	[3.639]
SFA	1.047	0.443	**1.098**	0.382	0.506	0.679
	[1.291]	[2.794]	**[1.295]**	[2.799]	[1.149]	[2.615]

a. Estimates are each drawn from separate regressions controlling for several student and school characteristics. See Appendix Table A1. Figures in brackets are robust standard errors

b. Heckman selection correction procedure used to account for fact that the sample of students used to estimate model impacts, namely students who have required tests scores and who remain in treatment or comparison group schools, is a non-random selection of all students originally in the treatment and comparison group schools. The selection term is significant for entries in bold.

c. Sample includes students with required test scores who moved out of a treatment or comparison group school to another New York City school. Students in a school that is not part of the treatment group or the original comparison group are counted as part of the comparison group.

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	IV	OLS	IV	OLS	IV
Basic Estimates						
SDP	1.750	1.118	-0.782	0.955	0.028	0.542
	[1.609]	[1.948]	[0.592]	[1.024]	[0.724]	[1.234]
MES			0.589	4.924**	-0.010	-0.477
			[0.865]	[1.973]	[0.650]	[1.742]
SFA			0.418	-0.032	-2.247**	-2.093
			[1.002]	[1.549]	[0.622]	[1.984]
With Heckman Selection Correction[b]						
SDP	**1.726**	**1.166**	-0.784	0.966	0.025	0.535
	[1.603]	**[1.942]**	[0.589]	[1.027]	[0.717]	[1.174]
MES			**0.632**	4.782**	-0.011	-0.425
			[0.850]	[1.961]	[0.631]	[1.752]
SFA			**0.471**	0.044	**-2.170****	-1.990
			[1.006]	[1.534]	**[0.610]**	[1.969]
Including Movers[c]						
SDP	1.750	1.118	-0.732	1.040	-0.078	0.955
	[1.609]	[1.948]	[0.584]	[1.033]	[0.651]	[1.127]
MES			0.673	4.858**	-0.017	-0.067
			[0.841]	[1.722]	[0.605]	[1.746]
SFA			0.388	-0.567	-2.213**	-3.916*
			[0.986]	[1.639]	[0.573]	[2.150]
With Measurement Error Correction[d]						
SDP					-0.483	-0.255
					[0.707]	[1.216]
MES					-0.454	-0.670
					[0.641]	[1.670]
SFA					-2.169**	-1.980
					[0.616]	[2.046]

a. Estimates are each drawn from separate regressions controlling for several student and school characteristics. See Appendix Table A1. Figures in brackets are robust standard errors.

b. Heckman selection correction procedure used to account for fact that the sample of students used to estimate model impacts, namely students who have required tests scores and who remain in treatment or comparison group schools, is a non-random selection of all students originally in the treatment and comparison group schools. The selection term is significant for entries in bold.

c. Sample includes students with required test scores who moved out of a treatment or comparison group school to another New York City school. Students in a school that is not part of the treatment groups or the original comparison group are counted as part of the comparison group.

d. Two-year test score lag is used as an instrument for the lagged-dependent variable, in order to correct for potential measurement error. A sufficient sample of observations with two-year test score lags is available only for fifth graders.

Table 7. Comparison of Methods for Estimating Impact of Whole-School Reform on 5th Grade Reading Scores

	D-VA	D-VA-IV	VA	VA-IV	LEV	LEV-IV
More Effective Schools	-0.494	1.512	0.261	2.902	0.074	2.612
	[0.97]	[1.51]	[0.91]	[2.12]	[1.25]	[2.38]
Success for All	-3.602**	-2.275**	-2.595**	-2.490**	-3.212**	-3.197*
	[1.09]	[1.02]	[0.77]	[0.96]	[1.61]	[1.76]
N	4,483	4,483	4,483	4,483	4,483	4,483
R-squared	0.092	0.089	0.581	0.579	0.143	0.141

Figures in brackets are robust standard errors; * = significant at the 0.10 level; ** = significant at the 0.05 level. D = difference-in-difference, VA = value-added, IV = instrumental variable, LEV = simple level regression. Regressions based on students in 1994-95 cohort who attended schools that implemented whole-school reform in 1995-96 or 1996-97 or appropriate comparison schools. Regressions include a full set of control variables, plus a Heckman selection correction. See Appendix Table A1.

TABLE 8. Variation in the Estimated Impact of Whole-School Reform by Student Characteristics[a]

		Cumulative Impact Through Grade 3		Value-Added Impact			
				Grade 4		Grade 5	
		OLS	IV	OLS	IV	OLS	IV
SDP	Not Free-Lunch Eligible	5.320*	7.275*	0.266	0.464	-1.019	-0.584
		[2.941]	[4.425]	[1.034]	[1.545]	[1.075]	[1.302]
	Free-Lunch Eligible	0.709	2.188	-0.900	1.001	0.146	0.662
		[1.377]	[2.158]	[0.590]	[1.073]	[0.734]	[1.293]
	Statistically Significant Difference[b]	YES	NO	NO	NO	NO	NO
	Not ESL Eligible	1.032	2.724	-0.820	1.055	-0.077	0.630
		[1.518]	[2.275]	[0.624]	[1.069]	[0.756]	[1.302]
	ESL Eligible	2.051	2.491	-0.514	0.089	0.803	-0.333
		[1.961]	[4.079]	[1.057]	[2.084]	[0.940]	[1.441]
	Statistically Significant Difference[b]	NO	NO	NO	NO	NO	NO
	Black	1.053	3.710	-0.811	1.666	-0.187	0.468
		[1.642]	[2.272]	[0.674]	[1.151]	[0.897]	[1.493]
	Hispanic	1.478	-0.240	-0.667	-0.978	0.524	0.839
		[1.493]	[2.753]	[0.799]	[1.323]	[0.666]	[1.105]
	Statistically Significant Difference[b]	NO	YES	NO	YES	NO	NO
MES	Not Free-Lunch Eligible	-4.055	-3.745	-2.127	4.845	-0.781	-0.040
		[2.633]	[2.868]	[1.985]	[3.248]	[1.139]	[5.729]
	Free-Lunch Eligible	4.175**	12.752**	0.679	4.992**	0.123	-0.562
		[1.243]	[3.405]	[0.894]	[2.026]	[0.660]	[1.671]
	Statistically Significant Difference[b]	YES	YES	NO	NO	NO	NO
	Not ESL Eligible	2.959*	15.947**	0.175	5.051**	-0.035	-0.235
		[1.561]	[4.726]	[1.068]	[2.528]	[0.737]	[1.899]
	ESL Eligible	2.554	4.685	0.736	3.696**	0.085	-1.411
		[1.993]	[2.984]	[1.119]	[1.747]	[1.113]	[2.234]
	Statistically Significant Difference[b]	NO	YES	NO	NO	NO	NO
	Black	2.600	7.14*	-1.239	5.359	-0.756	-1.960
		[2.349]	[3.703]	[1.432]	[3.438]	[0.936]	[3.358]
	Hispanic	3.327	14.334**	1.032	4.423**	0.421	-0.012
		[1.379]	[4.303]	[0.918]	[2.048]	[0.645]	[1.445]
	Statistically Significant Difference[b]	NO	NO	YES	NO	NO	NO
SFA	Not Free-Lunch Eligible	-0.804	-5.295	0.768	-1.040	-2.196**	-3.180**
		[2.463]	[5.998]	[1.718]	[3.210]	[0.942]	[1.367]
	Free-Lunch Eligible	1.178	1.077	0.382	0.057	-2.252**	-1.959
		[1.289]	[2.106]	[0.955]	[1.607]	[0.654]	[2.139]
	Statistically Significant Difference[b]	NO	NO	NO	NO	NO	NO
	Not ESL Eligible	0.795	0.834	0.566	0.053	-2.205**	-1.897
		[1.395]	[2.276]	[1.165]	[1.622]	[0.642]	[2.123]
	ESL Eligible	3.287**	-0.348	-0.633	-1.196	-2.524*	-3.776**
		[1.614]	[4.309]	[1.119]	[1.581]	[1.379]	[1.392]
	Statistically Significant Difference[b]	NO	NO	NO	NO	NO	NO
	Black	0.792	2.278	0.714	0.614	-2.706**	-2.184
		[1.647]	[2.897]	[1.105]	[1.831]	[0.702]	[2.742]
	Hispanic	1.919	-0.545	0.110	-1.375	-1.433*	-2.030**
		[1.356]	[2.677]	[0.928]	[1.521]	[0.778]	[0.977]
	Statistically Significant Difference[b]	NO	NO	NO	NO	NO	NO

a. Estimates are each drawn from separate regressions controlling for several student and school characteristics. See Appendix Table A1. Samples used are the same as in Table 6. Figures in brackets are robust standard errors; * = statistically significant at 0.10 level; ** = statistically significant at 0.05 level.

b. This row indicates whether the difference between the preceding two estimates is statistically significant at the 0.10 percent level or above.

TABLE 9. Variation in the Estimated Impact of Whole-School Reform by Quality of Model Implementation

	Cumulative Impact	Value Added Impacts	
	Through Grade 3	Grade 4	Grade 5
	OLS	OLS	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 Table A1. 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.

TABLE 10. Variation in the Impact of Whole-School Reform by Number of Years Implementing Reform Model[a]

	Implementing ≤ 2 Years		Implementing 3 or 4 Years	
	OLS	IV	OLS	IV
Cumulative Impact Through Grade 3[b]				
SDP	3.211**	6.816**	1.144	2.333
	[1.466]	[2.963]	[1.437]	[2.235]
MES	4.332**	19.695**	2.878**	12.939**
	[1.820]	[6.418]	[1.437]	[3.601]
SFA	-1.095	4.649	1.047	0.443
	[1.778]	[3.305]	[1.291]	[2.794]
N (SDP)	2,453		3,253	
N (MES)	1,037		855	
N (SFA)	931		885	
N (COMP)	8,319		5,685	
Value-Added Impact in Grade 4[c]				
SDP	-0.971	0.317	-0.173	1.012
	[0.702]	[1.321]	[0.878]	[1.315]
MES	1.478	3.781**	-0.994	6.038
	[0.800]	[1.536]	[1.274]	[3.927]
SFA	0.719	-0.588	1.330	1.766
	[0.702]	[1.440]	[1.412]	[1.971]
N (SDP)	1,771		1,579	
N (MES)	671		525	
N (SFA)	518		446	
N (COMP)	2,927		3,006	
Annual Value-Added Impact in Grade 5[d]				
SDP	-1.722		2.424	
	[2.022]		[2.403]	
MES	-0.012	-2.112	0.372	-0.811
	[0.878]	[2.636]	[0.743]	[2.279]
SFA	-3.715**	-2.164	-1.738	-0.107
	[0.958]	[2.306]	[1.077]	[2.589]
N (SDP)	149		152	
N (MES)	926		865	
N (SFA)	511		484	
N (COMP)	2,543		2,562	

a. Estimates are drawn from separate regressions controlling for several student and school characteristics. See Appendix Table A1. Figures in brackets are robust standard errors. N stands for number of observations; COMP indicates comparison groups.

b. Estimates in this panel are based on all treatment schools.

c. Estimates in this panel are based on schools that adopted in 1994-95 or 1995-96 which include 26 SDP, 6 MES and 6 SFA schools.

d. Estimates in this panel are based on schools that adopted in 1995-96 or 1996-97 which include 3 SDP, 10 MES and 7 SFA schools. There were not enough SDP schools to implement the IV estimation.

TABLE 11. Variation in Estimated Impact of Whole-School Reform by Number of Years Student Has Been Exposed

	SDP		MES		SFA	
	OLS	IV	OLS	IV	OLS	IV
Impact Through Grade 3						
One Year	1.901	2.778*	2.058	4.309	0.000	0.566
	[1.555]	[1.606]	[1.847]	[2.909]	[1.723]	[2.084]
Two Years	0.205	1.084	1.062	3.982*	-2.522	-2.338
	[1.349]	[1.939]	[1.460]	[2.205]	[2.274]	[2.408]
Three Years	0.749	2.552	3.818**	6.427**	0.475	0.799
	[1.502]	[2.910]	[1.428]	[1.583]	[1.669]	[3.303]
Four Years	1.121	1.335	1.706	15.857**	1.559	0.955
	[1.441]	[1.852]	[2.055]	[6.757]	[1.440]	[1.609]
N	11,776		8,698		8,768	
Annual Value-added in Grades 4 and 5						
One Or Two Years	-0.054	1.390	0.581	3.262	0.636	2.110
	[0.865]	[1.482]	[1.015]	[2.555]	[1.522]	[5.235]
Three Years	0.398	1.023	0.033	0.894	-1.364	-1.469
	[1.031]	[1.128]	[0.817]	[1.168]	[1.176]	[0.992]
Four Years	-0.905	0.043	0.253	1.421	-1.094	-0.852
	[1.046]	[0.839]	[0.832]	[1.091]	[1.184]	[1.262]
N	11,307		9,417		9,170	

Estimates are based on regressions controlling for several student and school characteristics. See Appendix Table A1. These regressions only include students either in a school that has implemented whole-school reform for three or four years or in a comparison school.

Table A1. 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	12,809	12,809	11,465	11,465	10,772	10,772
Uncensored Observations	8,261	8,261	6,899	6,899	6,393	6,393
I. The Production Function						
Adopted Whole-School Reform	0.025	0.535	-0.011	-0.425	-2.170**	-1.990
[Standard error]	[0.717]	[1.174]	[0.631]	[1.752]	[0.610]	[1.969]
Individual Characteristics						
Year=1999	-0.228	-0.241	-2.085	-2.041	-1.426	-1.384
Lagged Test-Score	0.643**	0.643**	0.627**	0.627**	0.629**	0.627**
Lagged Test-Score if >50	0.036**	0.035**	0.035**	0.035**	0.036**	0.038**
Lagged Test-Score*Year=1999	-0.009	-0.009	0.037	0.036	0.033	0.032
Lagged Test-Score if >50*Year=1999	-0.022	-0.022	-0.041**	-0.041**	-0.039**	-0.039**
Female	0.700**	0.711**	0.924	2.496**	0.967**	0.942**
Asian (reference category is white)	5.795**	5.106**	4.042	-25.392	1.371	1.100
Hispanic (reference category is white)	-0.996	-1.086	-1.154	-0.998	-0.996	-1.083
Black (reference category is white)	-2.279**	-2.251**	-2.301	-2.287**	-1.798**	-1.831*
Free Lunch Eligible	-1.390**	-1.273**	-1.395	2.797	-0.946	-0.869
Eligible for ESL Services[a]	-1.252	-1.598	-2.164	-17.170	-3.626**	-3.719**
Lamba (Inverse Mills Ratio)	-2.000	-0.823	0.264	63.027	6.396**	6.964
School Characteristics						
Log of Enrollment*10	0.136	0.153	0.306*	0.311**	0.159*	0.151*
% Free Lunch	0.023	0.025	0.039	0.042	0.063	0.060
% Limited English Proficient	0.023	0.016	-0.015	-0.009	0.028	0.024
% Hispanic	-0.023	-0.020	-0.007	-0.007	-0.016	-0.016
% Teachers <2 yrs experience	-0.061	-0.058	-0.008	-0.007	-0.003	-0.006
% Teachers w/certification	0.011	0.014	0.038	0.040	0.084*	0.079
Average Class-Size	-0.064	-0.071	-0.265*	-0.274*	-0.066	-0.050
SURR[b]	-0.655	-0.715	-0.982*	-0.884	-0.391	-0.476
II. The Selection Equation[c]						
Female	0.030		0.044**		0.031	
Asian	-0.822**		-0.722**		-0.567**	
Free Lunch Eligible	0.140		0.117		0.186**	
Eligible for ESL Services[a]	-0.361**		-0.397**		-0.335**	
Home-Language Other than English	-0.278**		0.005		-0.193**	

* 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 was 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.

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.

Bloom, Howard S. 1984. "Estimating the Effect of Job-training Programs, Using Longitudinal Data: Ashenfelter's Findings Reconsidered." *Journal of Human Resources* 19(4)(Fall): 544-556.

Bloom, Howard S., S. Ham, L. Melton, and L. O'Brien. 2001. *Evaluating the Accelerated Schools Approach: A Look at Early Implementation and Impacts on Student Achievement in Eight Elementary Schools.* New York: Manpower Demonstration Research Corporation.

Bound, J., D.A. Jaeger, and R.M. Baker. 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: 443-450.

Casserly, Michael. 2002. *Beating the Odds: A City-by-City Analysis of Student Performance and Achievement Gaps on State Assessments.* Washington, DC: Council of the Great City Schools, June. Accessed through http://www.cgcs.org/pdfs/beatodds2.pdf.

Comer, J.P., N.M. Haynes, and E.T. Joyner. 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., F. Habib, M. Phillips, R. Settersten, S.C. Shagle, and S.M. Degirmencioglu. 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., H.D. Hunt, and R.F. Murphy. 1998. "Comer's School Development Program in Chicago: A Theory-Based Evaluation." WP-98-24, Institute for Policy Research. Evanston, IL: Northwestern University.

Ferguson, Ronald, and Helen F. Ladd. 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.* Washington, DC: The Brookings Institution, pp. 265-298.

Goertz, Margaret E., and Mark C. Duffy, (with Kerstin Carlson 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. University of Pennsylvania Philadelphia, PA: March 2001.

Goertz, Margaret E., and Malik Edwards. 1999. "In Search of Excellence for All: The Courts and New Jersey School Finance Reform." *Journal of Education Finance* 25(1) (Summer): 5-31.

Heckman, J.J. 1979. "Sample Selection Bias As A Specification Error." *Econometrica* 47: 153-161.

Herman, R., D. Aladjam, P. McMahon, E. Masem, I. Mulligan, O. Smith, A. O'Malley, S. Quinones, A. Reeve, and D. Woodruff. 1999. *An Educator's Guide to School Wide Reform*. Arlington, VA: Educational Research Service.

Ladd, Helen F., and Janet S. Hansen. 1999. *Making Money Matter: Financing America's Schools*. Washington DC: National Academy Press.

Millsap, M.A., A. Chase, D. Obiedallah, and A. Perez-Smith. 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,* First Edition. Washington, DC: U.S. Department of Education.

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* CXIII, 553-602.

Schwartz, Amy Ellen, Leanna Stiefel, and Dae Yeop Kim. Forthcoming. "The Impact of School Reform on Student Performance: Evidence from the New York Network for School Renewal Project." *Journal of Human Resources*.

Slavin, R.E., N.A. Madden, L.J. Dolan, and B.A. Wasik. 1996. *Every Child, Every School: Success for All*. Newbury Park, CA: Corwin.

Viadero, Debra. 2001. "Memphis Scraps Redesign Models in All Its Schools." *Education Week* 20(42)(July 11): 1-19.

Wooldridge, Jeffrey M. 2003. *Introductory Econometrics: A Modern Approach*, 2nd Edition, Australia; Cincinnati, Ohio: Thompson, South-Western Publisher.