

**ADDRESSING SELF-SELECTION BIAS IN QUASI-EXPERIMENTAL
EVALUATIONS OF WHOLE-SCHOOL REFORM**

Robert Bifulco
Duke University

Center for Child and Family Policy
Duke University
Box 90264
Durham, NC 27708-0264
Office: (919) 613-7300
Home: (919) 383-3131
Fax: (919) 681-1533
bifulco@pps.duke.edu

AUTHOR'S NOTE: Funding provided by the Smith-Richardson Foundation helped to support this work. I would like to thank Bill Duncombe and John Yinger for their feedback on earlier drafts of this article. Please address inquiries to bifulco@pps.duke.edu.

Addressing Self-Selection Bias in Quasi-Experimental Evaluations of Whole-School Reform

Abstract

This paper discusses potential sources of self-selection bias in quasi-experimental evaluations of whole-school reform models, and considers how individual student level data might be used to provide valid impact estimates. While repeated pre- and post-treatment measures of student performance can provide unbiased estimates under relatively weak assumptions, such data are difficult to obtain. The paper develops an instrumental variable strategy that can be used to improve upon common value-added estimators when only post-treatment measures of performance are available. Using data from New York City, I show that the instrumental variable strategy can provide estimates of model impacts similar to those provided by a difference-in-differences estimator provided that appropriate instruments are used.

Keywords: whole-school reform, self-selection bias, impact analysis, quasi-experiment

Whole-school reform models offer replicable school management and instructional practices designed to improve student academic performance. In recent years, policy makers at the federal, state and district levels have been turning to these models to address concerns about low-levels of student achievement, particularly in schools with concentrations of disadvantaged students. Models that have figured prominently in recent policy initiatives include Success for All, School Development Program, Accelerated Schools, and Coalition of Essential Schools. Efforts to disseminate whole-school reform models represent a type of policy intervention that is common in education, but also in job training, welfare reform, and other social service fields. In this type of intervention the program participants are institutions (schools, job training offices, local welfare agencies, etc.), but the outcomes of interest are individual level variables.

As policy makers have begun turning to whole school reform models, efforts to provide rigorous estimates of model impacts have begun to emerge. Many of these emerging evaluations rely on quasi-experimental data. Quasi-experimental evaluations of institutional interventions, such as whole-school reform, must address two fundamental issues. The first issue concerns the unit of analysis. Should the evaluator take the institution as the unit of analyses and focus on estimating impacts on mean levels of individual outcomes across institutional units? Or should an evaluator take the individual as the unit of analyses, and focus on variation in outcome measures across individuals? The second issue concerns potential threats to internal validity posed by the fact that program participants are self-selected rather than randomly assigned.

Over the past decade, a consensus has begun to emerge favoring multilevel approaches to analyzing institutional interventions that explicitly embed individual-level models within institutional-level models. Such approaches can help researchers: formulate explicit hypotheses about the effect of institutional interventions on the relationship among individual-level variables; conduct appropriate statistical tests of these hypotheses; exploit the statistical precision provided by individual level variation without overestimating that precision; and avoid drawing fallacious inferences about individual level

processes from aggregate data. The problem is that many of the methods typically used to address potential self-selection biases in quasi-experimental analyses have not yet been adapted for use in estimating multilevel models. This creates a dilemma for evaluators trying to address both the unit-of-analysis and self-selection issues.

This article discusses the various forms of self-selection in quasi-experimental evaluations of whole-school reform models and examines alternative means of using multilevel data to provide valid impact estimates. I begin by presenting a hierarchical linear model (HLM) of student performance and identifying the assumptions that are required if maximum likelihood estimators are to provide valid estimates of this model. In Section 2, I explain why these assumptions are unlikely to be met in quasi-experimental evaluations of whole-school reform. In section 3, I discuss two alternative estimators. I argue that a difference-in-differences estimator that uses multiple pre- and post-test measures of student performance can provide valid impact estimates under relatively weak assumptions, but that the required data are difficult to obtain. Next, I describe an instrumental variable strategy that can be used to provide improved estimates in cases where only post-treatment measures of performance are available. Section 4, applies the HLM, difference-in-differences and instrumental variable estimators using data taken from an evaluation of whole-school reform in New York City, and compares the results. This empirical illustration suggests that commonly used HLM estimators might be biased, and that the use of appropriate instruments can remove part or all of this bias. The concluding section offers recommendations for emerging efforts to evaluate whole-school reform and other institutional interventions.

I. A Model of Program Outcomes

Different whole-school reform models posit various school and student outcomes as goals. The common goal across most, if not all, whole-school reform models is improved academic performance. Thus, I begin with the following hierarchical linear model of academic performance.

Student Level:

$$Y_{ijt} = \beta_{0j} + \beta_{1j}Y_{ij(t-1)} + \beta_{2j}X_{2ijt} + \beta_{3j}X_{3ijt} + \dots + \beta_{Qj}X_{Qijt} + r_{ijt}$$

$$r_{ijt} \sim N(0, \sigma^2) \quad \text{Cov}(X_{qijt}, r_{ijt}) = 0 \text{ for all } q$$

School Level: $\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1jt} + \gamma_{02}W_{2jt} + \dots + \gamma_{0S}W_{Sjt} + \delta T_{jt} + u_{jt}$

$$\beta_{qj} = \gamma_{q0} \quad q = (1, 2, 3, \dots, Q)$$

$$u_{jt} = N(0, \tau^2) \quad \text{Cov}(W_{sjt}, u_{jt}) = 0 \text{ for all } q \quad \text{Cov}(u_{jt}, r_{ijt}) = 0$$

Here Y represents the performance of student i in school j during year t ; each X_q is a student level variable that covaries with student performance, but is not itself influenced by the adoption of whole school reform; each W_s is a school-level measure and T is variable that indicates whether or not school j has adopted a particular whole-school reform model during or prior to year t . The model is a random intercept model in which the intercept in the student-level equation, β_{0j} , is assumed to vary as a stochastic function of various school-level factors. The model assumes that the effects of student-level variables on student performance are constant across schools. r_{ijt} and u_{jt} are random, student-level and school-level disturbances, both of which are assumed to be normally distributed with a zero mean and a constant variance. The model assumes that the student level error term is uncorrelated with the student level predictors and that the school level error term is uncorrelated with school level predictors. Finally it is assumed that r_{ijt} and u_{jt} are uncorrelated.

Combining the student-level and school-level models into a single equation yields:

$$(1) \quad Y_{ijt} = \gamma_{00} + \gamma_{10}Y_{ij(t-1)} + X'_{ijt}\Gamma_0 + W'_{jt}\Gamma_1 + \delta T_{jt} + u_{jt} + r_{ijt}$$

where X'_{ijt} is the vector of $X_{2ijt} \dots X_{Qijt}$, W'_{jt} is the vector of $W_{ijt} \dots W_{Sjt}$, and Γ_0 and Γ_1 are vectors of the parameters $\gamma_{20} \dots \gamma_{Q0}$ and $\gamma_{01} \dots \gamma_{0S}$, respectively.

Equation (1) is commonly used to investigate the determinants of student performance, and has two aspects that are worth noting. First, it includes a lagged measure of the dependent variable on the right-hand side. Inclusion of this measure reflects the cumulative nature of the education process, and is intended to capture the effect of past learning on a students' educational performance (Ferguson & Ladd 1996). Second, the random component of the model, $u_{jt} + r_{ijt}$, includes both a school-level and a student-level disturbance. This indicates that the combined error is not independently distributed across students, but rather is clustered by school. Failure to account for this clustering can bias estimates of standard

errors. Estimates of the parameters in this equation (γ_{kk} , δ , σ^2 , τ^2) can be obtained using maximum likelihood methods (Bryk & Raudenbush 1992).

The objective is to obtain estimates of δ that can be interpreted as the average impact of adopting the whole-school reform model, i.e. the average difference between a student's observed performance and what would have been observed if the school attended by that student had not adopted whole-school reform. Maximum likelihood will provide unbiased and consistent estimates of δ under the following conditions: the right-hand side variables in equation (1) are measured without error; the functional form of equation (1) is correct; and the treatment indicator, T , is uncorrelated with both school-level and student-level error terms, u_{jt} and r_{ijt} . Each of these conditions is potentially problematic.

Perhaps the most problematic form of measurement error in evaluations of whole-school reform is related to measurement of the intervention itself. Schools that decide to adopt a whole-school reform model vary in how well they implement the model. Moreover, the principles and practices associated with many models have diffused beyond the schools that have explicitly adopted whole-school reform. As a result, it is possible that an adopting school represents a particular model less truly than some non-adopting schools. This raises questions about how to define and measure the intervention represented by a whole-school reform model. In the estimations discussed below, the intervention is defined as the decision to adopt a particular whole-school reform model, which is measured as a simple dichotomy. Measurement error is less of an issue in this context than in evaluations that attempt to distinguish the quality of model implementation.

As written, the functional form of equation (1) assumes that the influences of individual characteristics on student performance and of school characteristics on the average level of performance in a school are linear. It also assumes that school level factors only influence the mean level of performance in a school, and not the relationship between individual characteristics and student performance. Together these assumptions imply that the impacts of whole-school reform do not systematically vary across either different types of schools or different types of students. This might not

be realistic. If model impacts vary across measured student or school characteristics, then these can be incorporated into equation (1) by using interaction terms between the treatment indicator and the relevant variables. Variation in model impacts across unobserved school or student characteristics can be more problematic. The implications of this unobserved heterogeneity in model impacts is discussed below.

The primary focus of this paper is on the potential correlation between the treatment indicator, T , and either the school-level or student-level error terms, u_j and r_{ijt} . Such correlation will bias estimates of whole-school reform impacts obtained from maximum likelihood estimates of equation (1). To assess the likelihood of this problem arising in quasi-experimental evaluations of whole-school reform, we need to consider the processes by which schools and students selected themselves into the intervention.

II. The Decision to Adopt a Whole-School Reform Model

Following a general approach to self-selection used in the economic evaluation literature, each school can be thought of as facing a set of expected benefits and costs resulting from the decision to adopt a whole-school reform model.¹ Benefits will include the expected gains in student performance following adoption. In addition, participation in a well-known model might provide school staff the opportunity to enhance their own human capital, gain favorable attention from colleagues and superiors, or otherwise advance career aspirations.² The costs of adoption include both pecuniary costs paid out of discretionary school funds and non-pecuniary costs related to the effort adoption would require. Pecuniary costs include payments for training and reimbursement to staff for participation in search, training and implementation activities. Non-pecuniary costs include those incurred during the search for a model, in achieving the consensus required to make the adoption decision,³ and during implementation.

The benefits and costs of adopting any given model will vary across schools. For instance, the School Development Program focuses attention on helping students from socially marginalized populations adjust to the social norms and demands of school. Thus, the expected gains from adoption of this model may be greatest for schools with high percentages of disadvantaged minority students. The availability of external funding will reduce the pecuniary costs to the school. Search costs will be determined largely by the availability of information on a model, and a school is more likely to have

information about innovations if the staff engages in more professional activities such as reading journals and attending conferences (Daft and Becker 1978). In cases where model developers demand a demonstration of staff consensus on the decision to adopt, the level of diversity or conflict among school staff might raise the costs of achieving consensus, and thus reduce the likelihood of adoption.

The presence of other schools in the district that have adopted a model, or support from the district office, may influence the costs of adoption in several ways. If a district agrees to pay for training or other implementation activities, then this reduces the amount the school needs to take from its discretionary budget. Economies of scale can be achieved by training more than one school at a time. Such collaboration is more likely if schools are from the same district. By providing information on a model, staff in other schools or in the central office familiar with a model can reduce search costs. Finally, district authority and personnel can help to build consensus prior to adoption and to facilitate implementation activities following adoption. In addition to reducing costs, a supportive district increases the “professional” or “bureaucratic” benefits associated with adoption.

This brief discussion of a school’s decision to adopt a whole-school reform model has two implications for attempts to estimate model impacts. First, many of the factors that affect the propensity to adopt will be unobserved. Some of these unobserved factors are likely to influence student performance as well. Formally, this implies that u_{jt} and T_{jt} in equation (1) are likely to be correlated. This correlation between treatment status and unobserved factors that influence student performance is the source of self-selection bias, and represents the primary problem for attempts to estimate the impacts of whole-school reform. Second, there are several potentially measurable variables, particularly district level variables, that provide sources of variation in a school’s decision to adopt that are arguably unrelated to student performance. As discussed below, these exogenous sources of variation in a school’s decision to adopt can provide means of addressing self-selection through the use of instrumental variables.

Before turning to the alternatives for addressing self-selection bias two additional points are worth making. This paper focuses on estimating the impacts of decision to adopt a whole-school reform

model, which is a dichotomous variable. In order to distinguish the impact of the decision to adopt a model from the impact of a well-implemented model, an analyst may want to use a multiple state treatment variable. For instance, a student can be in a school that has implemented the model well, one that has implemented poorly, or one that has not implemented at all. If unobserved factors that influence a school's capacity to implement a whole-school reform model are more strongly related to unobserved school quality than factors that influence the decision to adopt, then attempts to estimate the impact of the model when it is implemented well may be more susceptible to self-selection bias than the estimations presented below. Further, measurable variables that predict the decision to adopt might not predict the quality of implementation. Consequently, instrumental variable strategies that identify exogenous variation in the decision to adopt might not be useful for estimating the impact of well-implemented models.

It is also worth noting that our discussion of the school's decision to adopt whole-school reform ignores the potential influence of parental choices on program participation. The choices made by parents about where to send their children to school can give rise to differences between the students who attend schools with whole-school reform models and those who attend other schools. To see this, consider a case where schools are chosen to adopt whole-school reform through random assignment. In this case, we would expect the average characteristics of students in adopting schools to be the same as those in other schools at the time of adoption (Bloom, Bos, and Lee 1999). Nonetheless, differences between the students who attend adopting schools and those who attend other schools can emerge in ensuing years as students move into and out of the two sets of schools. If parents' decisions about where to send their children to school are responsive to the decisions of schools to adopt whole-school reform, then we might expect differences between the students in adopting and in non-adopting schools to emerge.⁴

Formally, parental decisions about where to send their children to school can generate correlation between r_{ijt} and T_j in equation (1), which would be an additional source of self-selection bias. Given the limited information parents have about whole-school reform models and the magnitude of other considerations that influence parents' decisions about where to send their children to school, the

likelihood of this type of bias in evaluations of whole-school reform models is low. Consequently, the discussions that follow largely ignore this issue. Nonetheless, strategies for dealing with this potential source of bias would be useful.

III. Alternative Estimators

Difference-in-Differences

Repeated measures for individual students can help to address self-selection bias. One way to exploit repeated measures of individual students is to construct a difference-in-differences estimator. Let Y_t^m and Y_{t-1}^m be the average performance of students attending schools that have adopted a whole-school reform model during two different years following adoption, $Y_{t^*}^m$ and $Y_{t^*-1}^m$ be two measures of performance prior to model adoption, where $t-(t-1) = t^*-(t^*-1)$, and $Y_t^c, Y_{t-1}^c, Y_{t^*}^c$ and $Y_{t^*-1}^c$ be the average performance of comparison group students during the same years. The difference-in-differences estimator is

$$\delta = \{(Y_t^m - Y_{t-1}^m) - (Y_{t^*}^m - Y_{t^*-1}^m)\} - \{(Y_t^c - Y_{t-1}^c) - (Y_{t^*}^c - Y_{t^*-1}^c)\}$$

More sophisticated methods adjust the comparison used to construct the difference-in-differences estimator for changes in observable factors that are unaffected by model adoption, but which might independently affect student performance. One way to implement this approach is by differencing equation (1):

$$(2) \quad Y_{ijt} - Y_{ijt^*} = \gamma_{00} + \gamma_{10}(Y_{ij(t-1)} - Y_{ij(t^*-1)}) + (X'_{ijt} - X'_{ijt^*})\Gamma_0 + (W'_{jt} - W'_{jt^*})\Gamma_1 \\ + \delta(T_{jt} - T_{jt^*}) + (u_{jt} - u_{jt^*}) + (r_{ijt} - r_{ijt^*})$$

Here t is a post-adoption year, $t-1$ is the year prior and is also a post-adoption year, t^* is a pre-adoption year, and (t^*-1) is the year prior to t^* . The maximum likelihood estimate of δ in this equation tells us the difference between the annual performance gains observed for those attending whole-school reform schools and the gains observed for the comparison group students controlling for the annual performance gains observed prior to the decision to adopt whole-school reform, and for any changes in observed student or school factors that are not influenced by whole-school reform. This estimate will be an

unbiased estimate of the impact of whole-school reform on student performance only if the right-hand side variables in equation (2) are measured without error; the functional form of equation (2) is correct; and both $(u_{jt}-u_{jt}^*)$ and $(r_{ijt}-r_{ijt}^*)$ are uncorrelated with treatment status.⁵

The assumption that $(u_{jt}-u_{jt}^*)$ in equation (2) is uncorrelated with treatment status is more plausible than the corresponding assumption that u_{jt} in equation (1) is uncorrelated with treatment status. The reason is that the effects of unobserved school characteristics on a school's decision to adopt a whole-school reform model, which are buried in u_{jt} , are likely to be more or less constant over time. Assuming a student has not changed schools, any time-invariant effects on student performance are differenced out of $(u_{jt}-u_{jt}^*)$ in equation (2). What are left in $(u_{jt}-u_{jt}^*)$ are changes in the effects of unobserved school characteristics on student performance. It is plausible to argue that these are either unrelated to the decision to adopt a whole-school reform model, or are themselves part of the changes caused by the decision to adopt whole-school reform.

The validity of the assumption that $(r_{ijt}-r_{ijt}^*)$ in equation (2) is uncorrelated with treatment status depends upon the growth trajectories that we expect students to follow as they move through elementary school. If annual growth rates of individual students tend to be constant as they move through elementary school, even if those rates differ across students, then there is little reason to think $(r_{ijt}-r_{ijt}^*)$ is correlated with treatment status. If, however, growth tends to either accelerate or decelerate as students move through schools, and the rate of acceleration varies systematically either across students or across schools, then the assumption may not hold. Since little is known about student growth trajectories, it is difficult to assess the plausibility of assuming equal acceleration (or deceleration) in growth rates across students.

Instrumental Variables

Difference-in-differences can provide defensible estimates of the impacts of whole-school reform. However, implementing this estimator requires at least two measures of student performance prior to the adoption of a whole-school reform model. Observers of whole-school reform argue that it may take several years before a whole-school reform model can be fully implemented and for improvements in student performance to be realized. Thus, the most interesting student cohorts to

examine in an evaluation of whole-school reform are those that are in the school several years after initiation of the reform. In the case of elementary schools, two years of student test scores prior to model adoption will not be available for these “most interesting” cohorts. Consequently, an alternative approach to addressing self-selection bias is often needed.

Instrumental variables (IV) estimators seek to overcome the self-selection problem by identifying a source of variation in who is exposed to a whole-school reform model that is unrelated to the unobserved variables that influence student performance. This requires a variable that meets two conditions. The first condition is that the variable provide an adequately precise prediction of whether or not school j attended by student i has adopted a whole-school reform model. The second condition is that the variable is uncorrelated with the unobserved factors that influence student performance.

The above discussion of a school’s choice to adopt a whole-school reform model suggests that district level variables might provide appropriate instruments. The presence of other adopting schools in the 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. Thus, we expect that a school will be more likely to adopt a given model, if other schools in the district have adopted the model. Whether the presence of other adopting schools in the district is uncorrelated with unobserved influences on student performance depends on the reasons why those other schools in the district adopted.

Consider the following:

$$(1) \quad Y_{ijt} = \gamma_{00} + \gamma_{10}Y_{ij(t-1)} + X'_{ijt}\Gamma_0 + W'_{jt}\Gamma_1 + \delta T_{jt} + u_{jt} + r_{ijt}$$

$$(3.J) \quad T_{jt} = f(Z_{1jt}, Z_{2jt}, \dots, Z_{Mjt}, T_{kt}, v_{jt})$$

$$(3.K) \quad T_{kt} = g(Z_{1kt}, Z_{2kt}, \dots, Z_{Mkt}, T_{jt}, v_{kt})$$

$$j \neq k \quad \text{cov}[u_{jt}, v_{jt}] \neq 0 \quad \text{cov}[v_{jt}, v_{kt}] \neq 0$$

Equations (3.J) and (3.K) predict the decision of j and k , respectively, to adopt a particular whole-school reform model, where j and k are different schools from the same district. Z_{mjt} ($m=1,2,\dots,M$) are measurable school level predictors and v_{jt} represents the influence of unobserved school characteristics on

the decision to adopt. Assume that the influence of unobserved variables on the schools decision to adopt (v_{jt}) is correlated with the influence of unobserved variables on student performance (u_{jt}). This assumption implies that T_j is correlated with the unobserved effects in equation (1), which causes maximum likelihood estimates of equation (1) to be biased and inconsistent.

Because schools in the same district may draw their students from similar populations and use a similar, district level hiring process, we might suspect that unobserved characteristics of students and teachers in schools j and k are correlated, i.e. $\text{cov}[v_{jt}, v_{kt}] \neq 0$. If the unobserved variables that influence school k 's decision to adopt also influence student performance and are shared with school j , then school k 's propensity to adopt a whole-school reform model will be correlated with student performance in school j , i.e. $\text{cov}[v_{kt}, u_{jt}] \neq 0$. This implies that the number of schools in the district that have adopted, may not be an exogenous source of variation in a school's decision to adopt.

If, however, the decision of school k is driven primarily by observed characteristics of the school, Z_{kt} , then these observed characteristics may provide suitable instruments. By supposition Z_{kt} will be determinants of school k 's propensity to adopt, and if school k 's decision to adopt influences school j 's decision, then Z_k will also provide good predictors of school j 's decision to adopt. It is also unlikely that observed characteristics of school k have any direct influence on student performance in school j .⁶

An instrumental variable estimator provides consistent estimates of the coefficients in equation (1), including δ (the impact of the decision to adopt a whole-school reform model), if: the right-hand side variables in equation (1) are measured without error; the functional form of equation (1) is correct; and the set of instruments used are correlated with T_{jt} and uncorrelated with the unobserved factors that influence student performance ($u_{jt} + r_{ijt}$).

Two things are worth noting about the instruments suggested here. First, these instruments isolate variation in a school's decision to adopt a whole-school reform model that is unrelated to unobserved *school* characteristics that influence student performance. Nonetheless, correlation between treatment status and unobserved *student* characteristics that influence student performance may arise if

parental choices about where to send their children to school are influenced by whole-school reform adoption decisions. Thus, IV estimators that use the instruments discussed here will provide consistent estimators of model impacts only if we assume that the adoption of whole-school reform models do not significantly influence parental decisions.

Second, IV estimators may not provide consistent estimates of model impacts if model impacts vary across schools based on unobservable characteristics. Particularly, if variables that are unobserved by the evaluator influence the impact of a model, and the school staff's knowledge of these unobservables influences its decision to adopt, instrumental variable estimators will not be consistent. Thus, the IV estimators considered here guarantee valid estimates of model impacts only if one of two conditions hold—either model impacts are constant across schools that are similar on observable variables, or if there is unobserved heterogeneity in impacts, then decisions to adopt whole-school reform are not influenced by it. The difference-in-differences estimator, in contrast, is robust to unobserved heterogeneity in model impacts (Heckman, Lalonde, and Smith 1999).

More generally, instrumental variable estimators can provide consistent estimates of model impacts, but these estimates may still be biased in finite samples. The magnitude of bias in finite samples depends on the sample size and the amount of variation in treatment status predicted or explained by the instruments. Bound, Jaeger, and Baker (1995) demonstrate that such bias can be quite serious when the instruments are weak predictors of treatment status. Thus, IV estimates of model impacts in finite samples tend to be sensitive to the choice of instruments, and if instruments are poorly correlated with treatment status, particular IV estimates can be quite misleading.

Finally, instrumental variable estimators do not account for the grouping of students within schools, and as a result provide biased standard error estimates. In the estimations below, I address this problem by computing robust standard error estimates, which allow for valid inferences in cases where errors are clustered by school.

IV. An Empirical Comparison of Alternative Estimators

In this section, I implement each of the three estimators discussed above using data from New York City. Based on the above discussion, I argue that the difference-in-differences estimates of model impacts are the most defensible. Assuming that the difference-in-differences estimates are unbiased, the results from this empirical exercise suggest that the hierarchical linear, value-added model may provide biased estimates of model impacts and that the use of appropriate instruments can help remove part or all of this bias.

Data

The data for this exercise are taken from an evaluation of whole-school reform efforts in New York City being conducted at Syracuse University's Center for Policy Research. The purpose of the study is to estimate the impacts of three different whole-school reform models (School Development Program, More Effective Schools, and Success for All) by comparing the performance of students in New York City elementary schools that adopted one of these models to the performance of students in a group of comparison schools.

The dataset assembled for the study includes individual student test histories, provided by the New York City Board of Education, for three cohorts of students. In particular, the data include results on citywide tests of math and reading for each student who attended third grade at one of the sample schools either in 1994-95, 1996-97 or 1998-99. The citywide tests were administered each year from either 2nd or 3rd grade (depending on the cohort) through 8th grade. Thus, the data include observations from multiple years for most students. Each student observation can be linked to school level data for the same year obtained from the New York City Board of Education's Annual School Reports and the New York State Education Department's Basic Education Data System.

The schools in the study that adopted whole-school reform did so during either the 1994-95, 1995-96 or 1996-97 school years. The majority of the students in the study sample who attended model adopting schools were exposed to whole-school reform prior to taking the citywide tests for the first time. Nonetheless, there is a subset of the treatment group students for whom two pre-exposure measures of

student performance are available. The cohort of students in third grade in 1994-95 who attended a school that adopted whole-school reform in either 1995-96 or 1996-97 have at least two years of pre-exposure test scores—namely their second and third grade test scores. Schools that adopted whole-school reform in 1995-96 or 1996-97 include 10 schools that adopted More Effective Schools (MES), 7 that adopted Success for All (SFA), and 3 that adopted the School Development Program (SDP). Because the number of schools adopting SDP is so small, the impact estimates are relatively imprecise and unstable. Thus, only students from MES and SFA adopters are included in these analyses.

In addition to these treatment group students, students who attended third grade during 1994-95 at a set of comparison group schools are included in the sample. The comparison group schools were selected using a stratified random sampling procedure from a set of New York City elementary schools that showed aggregate levels of performance similar to the treatment group schools in the three years preceding the 1995-96 school year or the three years preceding the 1996-97 school year. In total the comparison group sample includes 21 schools.⁷

The cohort of students in third grade during the 1994-95 school year in one of these 17 treatment group or 21 comparison group schools totals 4,173. However, the samples of students used for these analyses were limited in two ways. First, the most data intensive estimator examined in this section requires a test score for each year from 1993-94 through 1996-97 (i.e. from second through fifth grade). To avoid confounding differences in impact estimates due to the choice of estimator with those due to sample differences, any student who was missing a test score in any of these years was dropped from the sample. Second, the analyses here examine the performance of students during the 1996-97 school year. Many of the students who attended one of our sample schools in 1994-95 no longer did so in 1996-97. These students who were no longer in one of the treatment or comparison group schools during 1996-97 were dropped. The resulting sample used for the analyses of reading scores includes 2,070 students. In order to correct for potential biases that this sample selection might create, a Heckman selection correction term was estimated and included in the estimation procedures used here.⁸

Empirical Specifications

The three estimators discussed above (the value-added estimator, the difference-in-differences estimator, and the instrumental variable estimator) were implemented using these data. The outcome measure, treatment variables, and covariates used to specify the regression equations are detailed in Table 1. The outcome measure is the individual student's normal curve equivalent (NCE) score on the citywide test of reading. The New York City Board of Education changed reading tests in 1995-96 from the *Degrees of Reading Power* to the reading component of the *California Achievement Test-Series 5*. Because the NCE is a standardized test score, centered on 50, performance measures from these two different tests have the same interpretation and are commensurable. Nonetheless, we might expect a change in tests to affect test performance. The estimation procedures used here implicitly control for this change by including comparison group students who took the same tests in the same years as the treatment group students.

INSERT TABLE 1 ABOUT HERE

Two treatment variables are used, one for More Effective Schools and one for Success for All, each defined as a simple dichotomy. The covariates are self-explanatory except for the SURR variable indicating whether or not a school is under registration review. Registration review is a program administered by the New York State Education Department, which identifies schools as low-performing and requires identified schools to undertake specified improvement activities. Because schools under registration review (SURRs) were encouraged by the state education department to adopt a whole-school reform model, the whole-school reform adopters in the study sample are more likely to be SURRs than schools in the comparison group.

The validity of the estimators considered here requires correct specification of the functional form of the student performance equation. With two exceptions, each of the covariates listed in Table 1 are entered into the regression equation linearly. As indicated in Table 1, enrollment is entered into the

regression equation in log form, which was found to fit the data better. In addition, residual plots suggested that the lagged measure of student performance has a non-linear effect on the present year's student performance. In particular, students who score above average in the lagged year tend to show greater gains in the current year. To allow for this non-linearity, students with NCE scores above 50 in the lagged year were identified and the resulting indicator variable (=1 if NCE>50, =0 otherwise) is interacted with the lagged score. An extensive set of additional quadratic and interaction terms were entered into the equation both singly and in various combinations. In most cases these non-linear terms had statistically insignificant effects, and in the few cases where significant effects were found, these had insubstantial influence on the estimated impacts of the whole-school reform models.

The last variable in Table 1 requires comment. As discussed further below, several characteristics of other elementary schools in the community school district, excluding the school in which the student is enrolled, were tested as potential instruments for the decision to adopt a whole-school reform model. In the course of testing the appropriateness of these variables as instruments, it was discovered that the average percent of students eligible for free-lunch in the district had a significant, independent effect on student reading performance. One plausible explanation is that this measure is capturing a degree of concentrated poverty in the school that is not adequately captured by the school-level free lunch variable. In any case, the average percent of students eligible for free-lunch across other schools in the district is included as an additional school-level variable in the student performance equations.⁹

Estimation and Results

The results from each estimation procedure are presented in Table 2. The estimated impacts of MES and SFA are for students in the later elementary grades in schools that have been implementing whole-school reform for one or two years. The estimated coefficients on the MES and SFA variables indicate the average impact of the decision to adopt these models on student gains during the 1996-97 school year. These estimates miss any model impacts realized during the 1995-96 school year. In addition, whole-school reform developers and most observers would agree that whole-school reform can

take several years to begin showing positive impacts on student performance. Finally, these are estimates of the decision to adopt of model, and do not control for quality of implementation. For these reasons, conclusions about the efficacy of More Effective Schools and Success for All should not be drawn from the results presented here. Nevertheless, these analyses do serve to illustrate the methodological issues discussed above.

INSERT TABLE 2 ABOUT HERE

I begin by discussing the HLM-value added estimates presented in column one. This model was estimated using restricted maximum likelihood with the intercept term treated as a random effect. In addition, the diagnostic procedures recommended by Bryk & Raudenbush (1992) indicated that the effect of the lagged dependent variable measure should also be treated as random, and it is in the estimates presented here.¹⁰

Although we are primarily concerned with the estimated effects of MES and SFA several other results in column one deserve comment. The estimated coefficient on the lagged measure of student performance is highly significant. This estimate can be interpreted as the rate at which past learning decays over the time. The significant, positive coefficient on the lagged dependent variable for higher scoring students indicates that students who score well in one year retain more and/or gain more during the next year than do lower performing students. Among the other student level covariates the variable indicating whether or not a student repeated a grade has the largest impact. If we expect that repeaters are slower learners, the positive coefficient on this variable might seem perverse. For most students who have been retained, however, the lagged measure of performance is normed against the original cohort, while the current year performance measure is normed against a younger cohort. Thus, we expect a positive coefficient on this variable.¹¹ The inverse mills ratio (i.e. the Heckman selection correction) is also significant confirming the need to control for potential sample selection biases created by using only students with no missing test scores who remained in one of our sample schools.

Among the school-level variables, the percent of LEP students and the percent Hispanic students both have significant impacts in opposing directions. Students in schools currently under registration review show smaller performance gains. Whether this is due to negative effects of the registration review intervention, or to the effects of unobserved characteristics of schools under registration review is not clear. Finally, the negative effects of the percent of certified teachers, and positive effect of class-size are perverse. These last two results, which are robust across several specifications, are difficult to explain.

The decision to adopt MES shows a small, statistically insignificant, positive impact on student performance, while the decision to adopt SFA shows a larger, statistically significant, negative impact. The later result suggests that initial disruptions created by efforts to implement SFA, and possible diversions of school resources, have a negative effect on students in the later elementary school grades. However, because unobserved school factors that influence student performance gains are also expected to influence the decision to adopt whole-school reform, we suspect that the estimates in column one might be biased.

The second column of Table 2 presents difference-in-differences estimates. These estimates were obtained by subtracting the 1994-95 values of each of the variables in Table 1 from the 1996-97 values and using the differenced values to estimate the regression equation. In the case of the lagged dependent variable, the 1993-94 value is subtracted from the 1995-96 value. Again, restricted maximum likelihood was used to estimate the differenced equation. An intercept term was included in the estimated equation, and it was treated as a random effect. In this case, diagnostics revealed no reason to treat that effect of the lagged performance measure as random, and so it is treated as a fixed effect.

Differencing eliminates any variables that are constant over time, and thus many of the student level variables, including the Heckman selection term, drop out of the model in column two. In addition, much of the variation in school-level covariates is eliminated, and as a result, these variables have little influence in the model. The coefficient on the lagged dependent variable indicates the relationship of 1994-1996 (2nd grade to 4th grade) gains to 1995-97 (3rd grade to 5th grade) gains. Because the differences in gains realized from these overlapping periods are determined by differences between the 1994-1995

(3rd grade) gain and the 1996-97 (5th grade) gain, this coefficient is determined primarily by the relationship between student gains prior to model adoption and student gains following model adoption. The highly significant coefficient here indicates that pre-adoption gains are positive predictors of post-adoption gains. The negative coefficient on the interaction term immediately below the lagged performance measure indicates that students who scored below 50 in the 1994, but above 50 in 1996 showed smaller post-adoption gains than otherwise similar students. This might be explained by regression to the within-student mean. Finally, the variable indicating whether or not the student was retained shows even stronger positive impact than in column one. Because students retained are likely to have shown negative gains during the pre-test period (that is why they were retained), but positive gains in the year they are retained (when they are compared to younger students), this result is expected.

The difference-in-differences estimates in column two indicate that the decision to adopt More Effective Schools had a negative, but still small and statistically insignificant impact on student performance. The decision to adopt Success for All shows a negative impact which is similar to, although slightly larger than, that found using the value-added model. The difference-in-differences estimates in column two can be interpreted as the impact of MES and SFA on gains in student performance between 1996 and 1997, controlling for student gains made prior to model adoption and other changes in observable school characteristics. These are valid estimates of model impacts, if the effects of unobserved factors that influence both the decision to adopt whole-school reform and student performance are constant over time. This is more plausible than the assumption required by the value-added estimator that unobserved factors that influence student performance are unrelated to the decision to adopt whole-school reform. Thus, the estimates in column two are more defensible than those in column one. The difference between the two sets of estimates suggest that the estimates of model impacts obtained from the simple, value-added model do suffer from self-selection bias, although the bias in this sample appears to be minimal for SFA.

The third column of Table 2 attempts to improve upon the value-added estimates in column one by using two-stages least squares, which is an instrumental variables (IV) estimator. Drawing on the

earlier discussion of model selection, I focus on the characteristics of other schools in the same district as potential instruments. In selecting a set of instruments from among the several observed characteristics of other schools in district, two criteria were considered. First, the instruments chosen must be uncorrelated with the error term in the student performance equation ($u_{jt} + r_{ijt}$ in equation (1)). In cases where the number of instruments used is greater than the number of endogenous variables (in this case the MES and SFA indicators), it is possible to formally test for correlation between the instruments and the error term (Wooldridge 1999). Second, the instruments must be good predictors of the endogenous variables. In finite samples, IV estimates are biased in the direction of OLS estimates with the size of the bias depending on the strength of the relationship between the instruments and the endogenous variables. Bound, Jaeger, and Baker (1995) suggest that examining the F-statistic on the excluded instruments in the first stage regression is useful in gauging the bias of the IV estimator.

The instrument set used to generate the estimates presented in Table 2 includes the following measures from other schools in the same district: the log of the average enrollment; the average percent Hispanic students, the average percent of teachers with less than two years experience, the average percent of teachers who are certified in their field of assignment, and the square of the average percent of teachers certified. In choosing this instrument set, I first narrowed many different combinations of instruments to those for which the null hypotheses that the instruments are uncorrelated with the error term in the student performance equation could not be rejected. Among these sets of instruments, the one used has the highest partial F-statistic in the first stage regression. In the regression of these instruments on the MES indicator the partial F-statistic is 3.37 and the partial R^2 is 0.18. In the first stage regression on the SFA indicator the relevant F-statistic is 6.17 and the partial R^2 is 0.35. Because the errors in this model are clustered by school, I present robust standard error estimates calculated using the Huber-White procedure.¹²

Results for the control variables are similar to those obtained in column one. The estimated impacts of the decisions to adopt MES and SFA are, however, different. In particular, the IV-estimates in column three are closer to the difference-in-differences estimates than the value-added estimates in

column one. In the case of SFA, the IV and difference-in-differences estimates are virtually identical. However, the standard errors for the IV estimators are larger and thus the inferences differ. Whereas the estimated impacts of SFA are statistically significant at the 0.05 level in both columns one and two, they are significant only at the 0.10 level in column three.

It is important to note that the IV estimates are sensitive to the choice of instruments. For instance, suppose that we fail to recognize the independent relationship between student performance and the average percent of students eligible for free-lunch in the other schools in the district, and include this variable as an instrument in the first stage regression rather than as an independent variable in the second stage regression. In this case, we would reject the null hypothesis that the instruments are uncorrelated with error term in the student performance equation, and the impact estimates are markedly biased. Specifically, estimated coefficients for MES and SFA are -4.501 and -4.851 , respectively, in this misspecified model. Alternatively, suppose we drop the average percent Hispanic and the average percent new teachers from our set of instruments. Here, we do not reject the null hypothesis that the instruments are uncorrelated with the error term in the performance equation, but the relationships between this alternative set of instruments and the MES and SFA indicators is weaker than in the full set used to generate the estimates in Table 2. As a result, the estimated impacts obtained using this alternative set of instruments, 0.007 for MES and -3.388 for SFA, are closer to the value-added estimates in column one of Table 2 than are the IV-estimates presented in the third column of Table 2.

One concern with the estimates presented in Table 2 is that standardized tests are imperfect measures of student performance, even in the domains they are designed to assess. Thus, the lagged measures of student performance are measured with error. Although we expect that this error is randomly distributed across students, it nonetheless introduces bias and inconsistency into the coefficient estimates presented in Table 2 (Green 1997). To assess the extent of this bias, we reestimated the models in Table 2, using an instrument for the lagged performance measures in each model. In the value-added and IV models, the 1995 reading score was used as an instrument for the 1996 score. In the difference-in-differences model we used the 1994 reading score as an instrument for the difference between the 1996

and 1994 scores. If these instruments are uncorrelated with the error around the lagged measure of student performance (and since this error is randomly distributed they should be), and are good predictors of the lagged performance measure, then these alternative estimations will reduce the amount of bias due to measurement error.

The results of these alternative estimations are presented in Table 3. For each set of estimates, robust standard errors are used to account for clustering within schools. The point estimates do differ from the point estimates in Table 2. However, the qualitative pattern of results is the same. Assuming that the difference-in-differences estimates are our best estimates, and are unbiased, then there is reason to conclude that the value-added estimates are biased. In fact, the bias appears greater in Table 3 than in Table 2. Using instruments for the MES and SFA indicators, as well as for the lagged measure of student performance, reduces the bias of the value-added measures and provides impact estimates closer to the difference-in-differences estimates.¹³

INSERT TABLE 3 ABOUT HERE

V. Discussion

The discussion and empirical exercise presented above illustrate the difficulty of obtaining valid impact estimates for institutional interventions using quasi-experimental data. These difficulties are created by multiple potential sources of self-selection bias. In the case of whole-school reform, unobserved school factors influence a schools decision to adopt a model, other unobserved school factors affect the quality of model implementation, and unobserved student and family characteristics can influence which students attend whole-school reform schools. If any of these unobserved factors also influence student performance, estimated model impacts could be biased. Thus, even when individual-level performance data are available and modeled hierarchically, obtaining valid impact estimates presents challenges.

The preceding sections have considered three estimators that might be used to estimate the impact of the decision to adopt a whole-school reform model on student performance. We have seen that these estimators depend on several assumptions. What I have called the value-added estimator requires that unobserved factors that influence a schools decision to adopt a whole-school reform, or a students decision to attend a whole-school reform school, are unrelated to unobserved factors that influence student performance gains. The difference-in-differences estimator requires that the effects of unobserved factors that influence both the decision to adopt whole-school reform and student performance are constant over time. Instrumental variable estimators can provide consistent estimates if the instruments for the decision to adopt whole-school reform are uncorrelated with the unobserved factors that influence student performance, and unobserved factors do not influence the impact of the decision to adopt. In addition, each of the three estimators considered depend on the usual assumptions made in regression analysis that the independent variables in the regression equation are measured without error, that the functional form of the regression equation is correct, and that regressors are uncorrelated with the stochastic component of the model. Anyone of these assumptions might be difficult to test in a specific situation, and thus impact estimates are often accompanied by considerable uncertainty.

I have argued that the difference-in-differences estimator provides the most defensible estimates of model impacts. This is particularly true when techniques can be used to correct for random error in student performance measures. It is plausible to argue that changes in the effects of unobserved school characteristics on student performance are either unrelated to the decision to adopt a whole-school reform or are themselves part of the changes caused by the decision to adopt whole-school reform. Despite the plausibility of difference-in-differences estimates, their validity in any given situation remains uncertain. If student learning accelerates at different rates across different types of students, and this variation in acceleration rates is correlated with treatment status, then difference-in-differences estimates may be biased. Perhaps more important for the purposes of evaluating whole-school reform, the multiple

measures of student performance prior to model exposure required by the difference-in-differences estimator are often not available.

If we assume that the difference-in-differences estimates presented in Section IV are relatively unbiased, then the results of the empirical exercise in that section suggest two things. First, the value-added model commonly used to estimate the impact of educational interventions may indeed provide biased estimates. That value-added estimates may be biased is suggested by consideration of the factors that influence a schools decision to adopt a whole-school reform model, and is confirmed by differences between the value-added and difference-in-differences estimates in Tables 2 and 3. Second, use of appropriate instruments for the decision to adopt a whole-school reform model can help to reduce the bias of value-added estimates that is due to self-selection.

Several qualifications of these findings are needed. First, in the case examined here, the bias in the value-added estimates does not appear to be large and does not result in different qualitative inferences. Second, the results of instrumental variables estimation are sensitive to the choice of instruments. If the instruments used are only weak predictors of the decision to adopt whole-school reform and/or are correlated with the error term in the student performance equation, then IV estimates can be highly misleading. In practice, it might be difficult to specify an appropriate instrument set. In fact, in my own attempts to analyze the impacts of whole-school reform on student scores in mathematics for this study, I was unable to find an instrument set that was both uncorrelated with the error term in the student performance equation and a sufficiently strong predictor of the decision to adopt whole-school reform. As a result, I was only able to obtain highly misleading and/or very imprecise impact estimates using two-stage least squares. Finally, it should be noted that the results of the empirical exercise presented here assumes that the regressors in the student performance equation used are properly specified and adequately measured, assumptions which could not be fully tested.

A couple of lessons can be drawn for future attempts to estimate the impacts of whole-school reform and other institutional interventions. First, whenever possible multiple estimators should be used. If the results of different estimators are markedly different, then careful consideration of the assumptions

can suggest which are more plausible, but uncertainty will remain. If, on the other hand, different estimators converge to similar estimates than we can have more confidence in the evaluation results. Second, if two-staged least squares is used, it is important to formally test the assumption that the instruments used are uncorrelated with the error term in the second stage regression, and to verify that the excluded instruments are strong, independent predictors of the decision to adopt.

Despite the difficulties involved in using quasi-experimental data to estimate the impacts of institutional interventions, most evaluations will rely on quasi-experimental designs and such studies have the potential to provide valuable information. As work to evaluate whole-school reform and other interventions moves forward, continued attention to methodological issues is needed. In particular, strategies for addressing self-selection in studies that attempt to distinguish between the impacts of well-implemented and poorly implemented reforms, and for addressing the self-selection of students into schools, are needed. Also, alternative sources of potential instruments for the schools decision to adopt a whole-school reform model could help in instances where the instruments suggested here do not work well.

NOTES

1. For examples of selection models that assume that individuals choose to participate in a program based on comparison of the expected benefits and costs see Heckman, LaLonde, and Smith (1999). In the literature on organizational change, motivation-resource theory similarly attempts to model organizational decisions to adopt innovations as a benefit-cost calculation (Downs & Mohr 1976).
2. There is evidence that these opportunistic motives may be as important or more important than expected gains in student performance. In a well-known study of federally funded education programs, Berman and McLaughlin (1978) found that in many cases project adoption was motivated by the availability of funds rather than the possibility of change in educational practice, or that school managers saw program adoption as a low cost way to cope with bureaucratic or political pressures. In another study, Huberman and Miles (1984) found that school officials who adopted innovations were less interested in educational benefits than in “professional” or “bureaucratic” rewards.
3. Many models require some demonstrated level of staff commitment prior to adoption. For instance, Success for All requires approval by 80 percent of the school staff in a secret ballot.
4. If one is concerned with the impact of a whole-school reform model on the aggregate level of student performance in a school, then one would not want to control for changes in student population caused by the school’s decision to adopt. If changes in student population are driving the increase in aggregate student performance, and adoption of the model is driving changes in student population, then controlling for these changes would lead to underestimates of program impacts. However, if one is concerned with estimating the average impact of a whole-school reform model on the performance of individual students, then changes in school populations is a potential source of bias.
5. Ordinary least squares will also provide unbiased estimates of δ under these conditions. However, unless corrections are made for the grouping of errors within schools, estimates of the standard errors will be too small.
6. The observed variables in school k , Z_k , might be correlated with the unobserved characteristics of school k that influence both the decision to adopt and student performance. If so, Z_k might be correlated

with the error term in the student performance equation, i.e. equation (1). Thus, it is important wherever possible to test for correlation between the instruments and the error term in the student performance equation.

7. For details on the procedure use to select comparison schools see Bifulco (2001).

8. For details on how the Heckman selection correction term was estimated see Bifulco (2001).

9. One might add an additional level to the hierarchical model presented above, and treat the last variable in Table 1 as a district level regressor. However, since this variable measures the average percent eligible for free-lunch for *other* schools in the district, it does vary across schools in the same district.

10. Explicitly accounting for random effects is intended to allow for more efficient estimates and correct standard errors. Note, however, that this procedure does not adjust the standard errors for heteroscedasticity created by the Heckman selection term.

11. To see this, consider two students in fourth grade during 1996. Assume one of the students was retained in fourth grade. This student's 1997 score reflects his or her performance on the fourth grade test and is normed against other fourth graders. A five point test score gain for this student is not the same as a the five point gain for the other student who moved on to fifth grade, who took the fifth grade version of the test and whose score was normed against other fifth graders.

12. Unlike those in column one, these standard error estimates are robust to the heteroscedasticity created by including a Heckman selection correction term in the regression model.

13. Other covariates might be measured with error as well. Of particular concern are several of the school-level covariates, which are measured at the school-level, when the classroom level might be more appropriate. Efforts to find appropriate instruments for these variables were unsuccessful, and thus I cannot determine the influence of this source of potential measurement error.

REFERENCES

- Berman, P. and M.W. McLaughlin. 1978. *Federal programs supporting educational change, Vol. VIII: Implementing and sustaining innovations*. U.S. Department of Health, Education and Welfare. Office of Education, R-1589/8-HEW.
- Bifulco, R. 2001. *Do whole-school reform models boost student performance: Evidence from New York City*. Doctoral dissertation, Syracuse University.
- Bloom, H. S., J.M. Bos, and S. Lee. 1999. Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review* 23(4): 445-469.
- 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.
- Bryk, A.S. and S.W. Raudenbush. 1992. *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.
- Daft, R.L. and S.W. Becker. 1978. *The innovative organization: Innovation adoption in school organizations*. New York: Elsevier.
- Downs, G.W. and L.B. Mohr. 1976. Conceptual issues in the study of innovation. *Administrative Science Quarterly*, 21: 700-714.
- Ferguson, R. and H.F. Ladd. 1996. How and why money matters: An analysis of Alabama schools. In *Holding schools accountable: Performance-based reform in education* edited by H.F. Ladd. Washington, DC: The Brookings Institution.
- Green, W.H. 1997. *Econometric analysis, third edition*. Upper Saddle River, NJ: Prentice Hall.
- Heckman, J. J., R.J. LaLonde and J.A. Smith. 1999. The economics and econometrics of active labor market programs. In *Handbook of labor economics, Volume III* edited by O. Ashenfelter and D. Card. New York: Elsevier Science/North-Holland.

Huberman, M.A. and M.B. Miles. 1984. *Innovation up close: How school improvement works*. New York: Plenum Press.

Wooldridge, J.M. 1999. *Introductory econometrics: A modern approach*. Mason, OH: South-Western College Publishing.

About the Author: Robert Bifulco is an Assistant Professor at the University of Connecticut in the Department of Political Science. He is currently on leave as a Research Scholar for the Center for Child and Family Policy, which is housed in Duke University's Terry Sanford Institute of Public Policy.

Table 1: Definition and Summary Statistics for Variables Used in Model Estimations

Variable Name	Variable Definition	Mean (SD)		
		MES	SFA	Comparisons
Sample Size		577	396	1097
Performance Variables:				
1997 Reading NCE	Normal curve equivalent score on the 1997 citywide reading assessment	44.4 (14.9)	41.1 (14.9)	43.5 (15.5)
1996 Reading NCE	Normal curve equivalent score on the 1996 citywide reading assessment	45.8 (17.0)	44.5 (17.2)	44.5 (17.0)
1995 Reading NCE	Normal curve equivalent score on the 1995 citywide reading assessment	38.5 (19.3)	38.7 (19.6)	36.7 (19.0)
1994 Reading NCE	Normal curve equivalent score on the 1994 citywide reading assessment	43.0 (21.5)	43.6 (21.9)	42.6 (20.7)
Treatment Variables:				
MES	=1 if school had adopted More Effective Schools; =0 otherwise			
SFA	=1 if school has adopted Success for All; =0 otherwise			
Student Level Covariates:				
Sex	=1 if female; =0 if male	0.516 (0.500)	0.497 (0.501)	0.555 (0.497)
Hispanic	=1 if Hispanic; =0 otherwise	0.556 (0.497)	0.240 (0.428)	0.395 (0.489)
Free Lunch Eligible	=1 if eligible for free lunch in 1999; =0 otherwise	0.832 (0.374)	0.886 (0.318)	0.890 (0.313)
Non-English Home Lang.	=1 if home language is other than English; =0 otherwise	0.516 (0.500)	0.126 (0.333)	0.346 (0.476)
Behind Grade	=1 if student repeated a grade between 1994-95 and 1996-97; =0 otherwise	0.036 (0.187)	0.058 (0.234)	0.076 (0.265)
School Level Covariates:				
Log Enrollment*10	Log of the number of students enrolled multiplied by 10	69.0 (2.9)	67.6 (2.9)	69.2 (5.3)
%Free Lunch	Percent of students eligible for free lunch	92.9 (8.2)	95.3 (4.3)	94.8 (4.2)
%LEP	Percent of students classified as limited English proficient	34.9 (23.2)	18.7 (7.8)	24.5 (16.2)
% Hispanic	Percent of students who are Hispanic	64.5 (26.7)	34.7 (11.5)	49.6 (29.0)
%New	Percent of teachers with less than two years experience in education	15.0 (7.3)	11.1 (7.0)	16.4 (8.7)
%Certified	Percent of teachers certified to teach in their field of assignment	77.4 (12.8)	87.2 (7.9)	81.1 (10.8)
Class Size	Average class size	28.4 (1.6)	28.4 (2.2)	28.1 (2.5)
SURR	=1 if school is under registration review; =0 otherwise	7/10*	3/7*	7/21*
% Free Lunch (District)	Average percent eligible for free-lunch in other schools in community school district	89.6 (5.2)	86.3 (5.5)	84.7 (10.1)

* Figures represents number of schools under registration review/total number of schools.

Table 2: Estimated Impacts of Whole-School Reform Models on 1997 Reading Scores

	Value-Added (HLM)^a	Difference-in- Differences (HLM)^b	Value-Added (IV)^c
<i>Treatment Variables:</i>			
MES	0.886 (0.918)	-0.782 (1.575)	-0.152 (3.163)
SFA	-3.383** (0.931)	-3.598** (1.601)	-3.596* (2.058)
<i>Student Level Covariates:</i>			
Lagged reading score	0.623** (0.025)	0.225** (0.024)	0.620** (0.031)
Lagged reading score if > 50	0.039** (0.012)	-0.026** (0.013)	0.038** (0.017)
Sex	-0.136 (0.434)		-0.174 (0.512)
Hispanic	0.881 (0.682)		0.916 (0.778)
Free Lunch Eligible	-0.195 (0.678)		-0.237 (0.671)
Non-English Home Lang.	1.731** (0.873)		2.107 (1.718)
Behind Grade	6.075** (0.930)	13.156** (1.178)	5.769** (1.045)
Inverse Mills Ratio	-5.874** (1.514)		-6.514* (3.365)
<i>School Level Covariates:</i>			
Log Enrollment*10	0.118 (0.113)	-0.017 (0.041)	0.140 (0.116)
%Free Lunch	-0.083 (0.064)	0.114 (0.088)	-0.126 (0.081)
%LEP	0.077** (0.031)	0.048 (0.053)	0.075** (0.033)
% Hispanic	-0.094** (0.024)	0.059 (0.193)	-0.089** (0.021)
%New	-0.022 (0.042)	0.029 (0.077)	-0.006 (0.038)
%Certified	-0.078** (0.035)	0.034 (0.056)	-0.090** (0.033)
Class Size	0.453** (0.187)	-0.138 (0.279)	0.356* (0.207)
SURR	-2.299** (0.726)	0.496 (0.758)	-2.020** (0.800)
% Free Lunch (District)	-0.154** (0.049)	-0.429 (0.403)	-0.154** (0.052)

a. Intercept and lagged reading score treated as having random effects. Standard errors in parantheses.

b. Intercept treated as a random effect. Standard errors in parentheses.

c. MES and SFA are treated as endogenous. Robust standard errors reported in parentheses.

* Significant at 0.10 level ** Significant at 0.05 level

Table 3: Estimated Impacts of Whole-School Reform Models on 1997 Reading Scores with Measurement Error Correction^a

	Value-Added (IV)	Difference-in- Differences (IV)	Value-Added (IV)
Endogenous Variables	Lagged reading score	Lagged reading score	MES, SFA & Lagged reading score
<i>Treatment Variables:</i>			
MES	0.781 (0.941)	-1.384 (1.422)	0.320 (2.264)
SFA	-2.297** (0.850)	-4.234** (1.265)	-3.300 (2.268)
<i>Student Level Covariates:</i>			
Lagged reading score	1.236** (0.060)	0.629** (0.053)	1.227** (0.058)
Lagged reading score if > 50	-0.231** (0.028)	-0.176** (0.017)	-0.227** (0.028)
Sex	-0.907* (0.500)		-0.991 (0.592)
Hispanic	1.430 (0.902)		1.518 (0.899)
Free Lunch Eligible	0.133 (0.787)		0.184 (0.829)
Non-English Home Lang.	0.319 (1.272)		0.579 (1.805)
Behind Grade	12.286** (1.303)	12.566** (1.374)	12.173** (1.301)
Inverse Mills Ratio	-1.270 (2.445)		-2.280 (4.004)
<i>School Level Covariates:</i>			
Log Enrollment*10	0.034 (0.087)	-0.063 (0.079)	0.021 (0.115)
%Free Lunch	-0.143** (0.050)	0.087 (0.069)	-0.151* (0.081)
%LEP	0.071** (0.035)	0.043 (0.048)	0.075** (0.036)
% Hispanic	-0.076 (0.023)	0.038 (0.157)	-0.081** (0.027)
%New	0.016 (0.042)	0.077 (0.061)	0.008 (0.048)
%Certified	-0.084** (0.034)	0.013 (0.064)	-0.082** (0.035)
Class Size	0.435** (0.160)	-0.147 (0.200)	0.476** (0.204)
SURR	-1.537** (0.608)	0.019 (1.090)	-1.438* (0.852)
% Free Lunch (District)	-0.135** (0.028)	-0.232 (0.337)	-0.136** (0.049)

a. Robust standard errors in parantheses.

* Significant at 0.10 level ** Significant at 0.05 level