Program on Education Policy and Governance Working Papers Series

PEPG 09-02

Impact of For-Profit and Nonprofit Management on Student Achievement: The Philadelphia Intervention, 2002-2008

Paul E. Peterson and Matthew M. Chingos

Program on Education Policy and Governance
Taubman Center on State and Local Government
Harvard's Kennedy School of Government
Harvard University

79 John F. Kennedy Street, Taubman 304
Cambridge, MA 02138
Tel: 617-495-7976/8312
Fax: 617-496-4428
www.ksg.harvard.edu/pepg/

Paul E. Peterson is the Henry Lee Shattuck Professor of Government and Director of the Program on Education Policy and Governance in the Kennedy School of Government. He is a Senior Fellow at the Hoover Institution, Stanford University

Matthew Chingos is a Research Fellow in the Program on Education Policy and Governance.

Impact of For-Profit and Nonprofit Management on Student Achievement: The Philadelphia Intervention, 2002-2008

Paul E. Peterson and Matthew M. Chingos

(Executive Summary)

The School District of Philadelphia, in the summer of 2002, at the request of the State of Pennsylvania, asked for-profit and nonprofit managers to participate in a substantial restructuring of its lowest-performing schools under the overall direction of the Philadelphia School Reform Commission (SRC). Thirty elementary and middle schools were contracted out to for-profit management organizations; 16 schools were contracted out to nonprofit organizations.

Using individual student test-score data made available by the School District of Philadelphia, we estimated the impact of for-profit and nonprofit management on student achievement by tracking the performance of students in math and reading from 2001 to 2008. The first two years (2001 and 2002) provide us with information about student performance prior to the management intervention, while the subsequent six years (2003, 2004, 2005, 2006, 2007 and 2008) provide information about student performance after the interventions had begun. The measures of test performance are taken from three tests: the Terra Nova, the Stanford 9, and the Pennsylvania State System of Assessment (PSSA), the test Pennsylvania administers to comply with the accountability requirements in the federal No Child Left Behind Act (NCLB).

Of the 30 schools included in the study that were under for-profit management, 20 were managed by Edison Schools, five by Victory Schools, and five by Chancellor Beacon Academies. Of the 16 schools included in the study that were managed by nonprofits, five were managed by Foundations, three by the University of Pennsylvania, five by Temple University, and three by Universal Companies. We compare the performance of the privately managed schools to that of 71 schools that remained under regular school district management whose students performed below the district median.

This paper includes information for two more years (school years ending in 2007 and 2008) beyond what was previously reported in Peterson and Chingos (2007). It also examines the impact on test scores of the five schools for whom the for-profit contract was revoked by the school district.

Methodology

Since the assigned schools were the lowest performing schools in the district, our preferred model for estimating impacts on student performance is a "fixed effects model" that takes into account unobserved student characteristics that do not vary over time. That model is used to estimate impacts on all students for whom at least two test-score gains are available, the minimum necessary to carry out a fixed effects analysis. Two additional models that control for observable (but not unobservable) characteristics are used to estimate effects for all those for whom at least one test-score gain was available. Informal adjustments to the preferred model are made to estimate impacts on the larger number of students for whom two or more year's worth of test scores were available.

Impacts of the interventions are calculated in standard deviations. On average, the test performance of a student in grades 2 through 8 in the Philadelphia school district improved by 1.0 standard deviations in math over the course of 2.4 years. In reading a gain of 1.0 standard deviations in reading performance took an average of 3.6 years. We use this information to convert standard deviation units into approximate years of learning.

Despite the large number of student observations, they were clustered in a small number of schools, making it possible to detect only large impacts at conventional levels of statistical significance.

Nonprofit Management

The impact of nonprofit management appears to have been negative. At schools under nonprofit management, students learned, on average for the six years, 21 percent of a standard deviation *less* in math each year than they would have had their school remained under regular district management. Calculated in terms of years of schooling, the negative impacts on math performance were, on average, approximately 50 percent of a year's worth of learning annually, a large impact. However, the negative impact was statistically significant in only the first year after the intervention began. In reading the average adverse impact of nonprofit management was roughly 10 percent of a standard deviation less annually, about 32 percent of a year's worth of learning each year. However, the effect on reading performance was statistically significant in only the first year after the intervention began.

The other two models suggest that impacts may have been somewhat different for those with fewer test scores. Adjusting for that possibility, an informal adjustment to the estimation yields an adverse impact of nonprofit management on math performance of 18 percent of a standard deviation. In reading, it is 14 percent of a standard deviation.

For-profit Management

The impact of for-profit management was generally positive, though only the math impacts are statistically significant. At schools under for-profit management, students learned in math, on average, 25 percent of a standard deviation more each year of the six years of the intervention than they would have had the school been under district management. The estimated impact each year was roughly 60 percent of a year's worth of learning, a large, statistically significant impact.

The other two models suggest that impacts may have been somewhat different for those with fewer test scores. Informally adjusting for that possibility, an alternative estimate of the positive impact of for profit management on math performance is 12 percent of a standard deviation, 29 percent of a year's worth of learning.

The estimated average annual impact on reading performance of for-profit management relative to district management is a positive 10 percent of a standard deviation, approximately 36 percent of a year's worth of reading. However, that impact is not statistically significant. When one informally adjusts for the possibility that the impacts are different for those with only two test scores, then the impact was just 4 percent of a standard deviation, about 14 percent of a year's worth of learning.

Differences between For-profit and Nonprofit Schools

The differential impact of for-profit and nonprofit management is especially sizeable. Using the estimates given above, students in schools under for-profit management were learning the equivalent of somewhere between 70 percent of a year to more than a full year's worth of learning in math each year than they would have had the schools been under nonprofit management. All estimations are statistically significant. In reading students learned approximately two-thirds of a year more in a for-profit school than they would had the school been under nonprofit management. All but one estimation are statistically significant.

In the following table the average of the annual estimated impacts of nonprofit and for-profit schools relative to district management is presented for each of the six years of the intervention in standard deviations and years worth of learning. Differential impacts of the nonprofit and for-profit schools are also provided. Estimates are provided from the "fixed effects" model (Model III) as well as those from an informal adjustment that takes into account a larger set of the student population.

Table Ex. Sum. 1: Summary of Estimated Impacts

	Math		**Reading**	
	Standard Deviations	**Years of Learning**	**Standard Deviations**	**Years of Learning**
Nonprofit*				
"Fixed effects" (Model III)	-0.21	-50%	-0.09	-32%
Adjusted Estimates	-0.18	-43%	-0.14	-50%
For-profits*				
"Fixed effects" (Model III)	0.25	60%	0.10	36%
Adjusted Estimates	0.12	29%	0.04	14%
Differential Impact**				
"Fixed effects (Model III)	0.46	110%	0.19	68%
Adjusted Estimates	0.30	72%	0.18	64%

*Impact relative to School District Management

**Impact of for-profit management relative to nonprofit management

Closing For-profit Schools

Ironically, the School District of Philadelphia reassumed the management of five of its for-profit schools but only one of its nonprofit schools. To ascertain whether that decision had a strong educational basis in the district's test score data base, we used the same methodology to estimate the impact on student learning of the five schools for whom the for-profit management contract had been terminated.

The results are mixed but provide little support for the district's decisions. On the one hand, the reading performance of students at the five schools under for-profit management was, on average, 18 percent of a standard deviation below what could have been expected had the schools been under district management, a difference that is statistically significant in three of the six years. Also, nonprofit schools whose contract was not revoked had an impact on reading performance that was 10 percent of a standard deviation more positive than that of the for-profit schools whose contract was terminated. However, that difference is not statistically significant in any year.

On the other hand, the math performance of students at the for-profit schools was 35 percent of a standard deviation higher than would have been the case had the schools been under district management. Also, it was 56 percent of a standard deviation higher than would have been the case had the schools been under nonprofit management. Those large differences are statistically significant in most years.

A case for terminating the management of the for-profit schools could be made on the basis of student test score performance only if the strongly positive math performances are ignored, as they clearly outweigh any adverse impacts in reading. However, the test score information suggests a case can be made for terminating the contracts of the nonprofit providers.

Impact of For-Profit and Nonprofit Management on Student Achievement: The Philadelphia Intervention, 2002-2008

Paul E. Peterson and Matthew M. Chingos

Abstract

At the request of the State of Pennsylvania, the School District of Philadelphia, in the summer of 2002, asked three for-profit firms to assume responsibility for 30 of its lowest-performing schools and it asked four nonprofit managers to assume the management of 16 other low-performing schools. A difference-in-differences analysis is used to estimate the impact of nonprofit and for-profit management on individual student achievement. Gains in test scores at the treated schools are estimated by comparing them with gains in other low-performing schools in the district.

Students at schools under for-profit management outperformed those at schools under nonprofit management in all six years in both reading and math. Most estimations are statistically significant. Impacts of for-profit management relative to district management were positive in math, but no reading impacts could be detected. At nonprofits, students appear to have learned substantially less, especially in math, at nonprofit schools, than had their school remained under regular district management. However, impacts fell short of statistical significance.

The U.S. federal law No Child Left Behind (NCLB) requires states to "restructure" any school that fails for six years running to make Adequate Yearly Progress (AYP) toward full proficiency on the part of all students by the year 2014.[1] The law provides a number of restructuring options, including the shift of the school's management to a private entity, either a for-profit firm or a nonprofit organization. Although only a few school districts nationwide have sought help from either for-profit or nonprofit organizations in the management of low-performing schools (Mead 2007), in 2002 the School District of Philadelphia, at the request of the State of Pennsylvania, asked both types of entities to participate in a substantial restructuring of many of its lowest-performing schools under the overall direction of the Philadelphia School Reform Commission (SRC). Thirty elementary and middle schools were contracted out to for-profit management organizations, and sixteen schools were contracted out to nonprofit organizations.[2]

The policy intervention in Philadelphia raises questions of general interest: Will students at schools assigned to for-profit or nonprofit managers learn more than would be expected had those schools remained under school district management? Is for-profit management more or less effective at raising achievement than nonprofit management?

The distinction between for-profit and nonprofit management has been a topic of continuing discussion in the scholarly literature on school reform. Friedman (1955) and Coulson (1999, pp. 304-306) have theorized that for-profit firms are more effective because they have clear economic incentives to lift student performance. The firm can build its reputation (and in the long run generate a profit) only if it becomes known for running effective schools. Others have suggested, however, that for-profit firms are likely to cut costs and shortchange students in order to benefit the firm's owners and shareholders (Hochschild and Scovronick 2003, pp. 120-21; Levin 2001). The debate over nonprofit organizations takes a different form. Some have argued that nonprofit managers are likely to be effective because they have close ties to the community within which they are embedded and can enlist the energies of committed entrepreneurs, who devote all available resources to enhance student performance (Brandl 1998; 2006). But others caution that nonprofit managers may have neither the experience, resources, nor economic incentives necessary for building quality educational institutions (Hassel 2003, pp. 190-93; Merrifield 2001, pp. 32-35).

Evidence concerning the impact of for-profit and nonprofit management has been anecdotal. To the best of our knowledge, no systematic evaluations of individual test score gains have estimated relative impacts under similar operating conditions.[3] In this paper, we provide empirical information about the impact of for-profit and nonprofit managers on student achievement in Philadelphia. This paper includes information for two more years (school years ending in 2007 and 2008) beyond what was previously reported in Peterson and Chingos (2007).

The Intervention

Only after an intense political struggle did the Philadelphia school district ask for-profit and nonprofit providers to assume responsibility for a number of its schools. In 2001 Pennsylvania's Republican Governor Tom Ridge, a school voucher supporter, indicated he would not support any increment in funding for the Philadelphia school district until an independent entity had assessed its financial practices and educational effectiveness. Philadelphia's mayor, Democrat John Street, knowing the district was facing a $215 million deficit, agreed to the study, and the state department of education asked Edison Schools to carry out the investigation. Edison, a for-profit firm that manages charter schools and other schools under contracts with school districts, reported that the Philadelphia school district had spent $10 billion "with no clear accountability for the results" (Matthews 2003, p. 54). Governor Ridge refused to give any more state aid beyond current levels unless the school district agreed to a new partnership with the state.

When it was reported in the local press that Edison was expected to assume management of many of the city's schools, the local teachers union mobilized in opposition, and groups of parent and student activists held rallies throughout the city. "From the moment I read about Edison and its history," said one parental opponent, "I was determined to keep them out of the Philadelphia schools" (Matthews 2003, p. 56). As the turmoil was reaching its climax, Governor Ridge resigned office to become the nation's Secretary of Homeland Security, and Pennsylvania's Lieutenant Governor, Mark Schweiker, was elevated to the state's highest office.

Upon assuming office, he said he would not serve beyond his current term ending in 2002, a decision that weakened his leverage vis-à-vis the Philadelphia school district.

The stage was set for a compromise that would save face for all the parties involved. It was agreed that the school board should be replaced by a School Reform Commission (SRC), three members of which would be appointed by the governor and two by the mayor. The SRC decided that only a limited number of the lowest-performing schools in the district would be turned over to private management. Edison Schools was not to be the only private provider. Instead, seven entities—three for-profit firms and four nonprofit entities—were chosen. The SRC explained its decision by saying that multiple providers would yield information on the kind of management that was most effective. Edison Schools was asked to manage 20 of the schools, with another 5 each to become the responsibility of Victory Schools and Chancellor Beacon Academies, two other for-profit companies. Another 16 of the low-performing schools would be managed by nonprofit entities— the University of Pennsylvania (3 schools), Temple University (5 schools), Foundations (5 schools), and Universal (3 schools). The SRC also appointed a reform-minded superintendent—energetic, outspoken Paul Vallas, who had instituted a series of reforms in Chicago at the behest of Mayor Richard Daley.

The for-profit firms had fewer local political connections than did the nonprofit entities. None of the for-profit companies had operated schools or other programs within the city. Though Edison was held in high regard by the Republican state secretary of education, it faced strong opposition within Philadelphia, especially from the Philadelphia Teachers Association. However, Edison Schools could claim considerable experience at running schools, as it was the manager of one hundred public and charter schools nationwide. Victory Schools, a company that offered single-sex education within classrooms, was the manager of schools in the state of New

York and in Baltimore, Maryland. To strengthen its local connections, it hired a former district employee to head up its Philadelphia effort. Chancellor Beacon operated some eighty private and charter schools, but it had not previously managed district-operated schools. Shortly after the intervention began, Superintendent Vallas cancelled the district's contract with Chancellor Beacon, and its five schools were either brought back under district control or assigned to other providers. (Following Gill et al. [2007], we assign them to treatment status in our analysis. However, results do not change materially if Chancellor Beacon schools are excluded.)[4]

By contrast, the nonprofit entities were—and have remained—politically well-connected institutions. The University of Pennsylvania is a Philadelphia icon, a highly prestigious Ivy League institution with a history dating back to Benjamin Franklin. Temple University's status is less exalted, but it is nonetheless a well-established Philadelphia institution of higher learning. Foundations was established by one of the school district's former associate superintendents, and "many of its staff members are former employees of the district" (Christman et al. 2006, p. 10). It has had close ties to a politically influential state representative who was active in community development programs. Universal, a community development corporation founded by Kenny Gamble, an immensely successful writer of soul music, has strong ties to Islamic leaders within Philadelphia's black community ("Gamble and Huff," 2008).

Temple University and the University of Pennsylvania drew upon the resources of their schools of education. Rather than taking on a general reorganization of the schools, they focused mainly on providing professional development to teachers, formal and informal assessment feedback to teachers, and within-classroom coaching services to students. Foundations operated after-school programs and favored a teaching approach that relied upon computer-based learning where students progressed at their own pace. Universal was known for "boosting economic development" and "providing social services," but it had only "limited school management experience" (Christman et al. 2006, p. 10).

The school district restricted the prerogatives of both for-profit and nonprofit managers in a number of ways. Management had to operate within the framework of the district's collective bargaining agreements with its union employees, and teachers were allowed to transfer to other schools within the district, if they wished. Many of them chose to do so—as could be expected, given the fact that the schools in question were considered to be those most in need of new management. The district also retained control over such areas as "facilities management, school safety, food services, the overall school calendar, decisions about holiday closures, altering grade configurations, and the code of conduct for teachers and students" (Christman et al. 2006, p. 11).

By 2008, several leadership changes had occurred. The SRC's new chair, Sandra Dungee Glenn, had been appointed by the Democratic Governor Edward Rendell (who had succeeded Schweiker). She was a former community organizer active in Democratic politics. After fighting with the SRC over a surprise budget deficit, Superintendent Paul Vallas left for New Orleans where he took on the daunting task of rebuilding the city's post-Katrina school system. The SRC hired a new superintendent, Arlene Ackerman, who had previously served as the superintendent of the school systems in San Francisco and Washington, DC.

Under its new leadership, the SRC removed five schools from the management of the for-profit firms (four from Edison, one from Victory) and one school was removed from the management of a nonprofit entity (Temple). Acting upon the superintendent's recommendation, the SRC decided that the six schools should come under direct district control once again, because they had not made Adequate Yearly Progress (AYP), as required by No Child Left Behind. "It's been six years—it is time to sort it out," said the SRC chair, Sandra Dungee Glenn

(Dean, 2008). Ackerman indicated that many more schools could come back under district control once a full-scale review had been undertaken.

The selection of schools to be taken back under district control could have been due to either educational considerations or other factors. On the one hand, the nonprofit schools were well-connected politically, while the for-profit firms were not. On the other hand, the nonprofits might have been the more effective educational institutions. To see whether this was the case, we use a quasi-experimental research design to identify the impact of for-profit and nonprofit management in Philadelphia.

The Data

Our analysis is based upon information supplied by the Philadelphia school district. Test score, demographic, and school enrollment information on Philadelphia students in grades 2-8 from 2001 through 2008 are available for each student.[5] The test score data come from three different tests (see Table 1). The Pennsylvania State System of Assessment (PSSA) is the test currently used for holding schools accountable for improved student learning in Philadelphia. But when the private management intervention began in the fall of 2002, that system of measuring school performance was still a work in progress. Only students in 5th, 8th, and 11th grades were given the PSSA test in reading and math. Not until the spring of 2005 did schools begin testing students with the PSSA in grade 3, and not until 2006 did the 3rd-grade exam become a part of the state accountability system. Grades 4, 6, and 7 were not tested until 2006. However, two other tests were given to some Philadelphia students between 2001 and 2006. In 2001 and 2002, students in grades 3, 4, and 7 were tested on the Stanford 9; between 2003 and 2005, the Terra Nova, designed by a different company, was used at most grade levels. The Stanford 9 and Terra Nova are nationally normed tests. Some students, in some years, were given more than one test. In 2006, the Terra Nova was dropped, except for students in first and second grade. The PSSA was given to 3rd graders in 2005, but it did not become a test that was used for accountability purposes until 2006, the same year the PSSA was introduced in grades 4, 6, and 7.[6] In order to make maximum use of these data, we standardize scores by test, grade, and year to have a mean of zero and standard deviation of one. For students who took two tests between 2003 and 2005 (many 5th and 8th graders took both the PSSA and Terra Nova from 2003 to 2005) we make use of Terra Nova scores because that test was used more consistently during that time period. For the 3rd through 8th graders who took both the PSSA and Terra Nova in 2007, we use the PSSA scores because that test was used almost exclusively in 2006 through 2008.

Students are matched to their respective schools, and schools are classified as either under for-profit management, under nonprofit management, or under regular district management.[7]

The average combined reading and math test scores prior to treatment at schools assigned to for-profit and nonprofit managers were 0.39 and 0.13 standard deviations below the Philadelphia average, respectively, while the pre-treatment scores of the full set of 142 regular public schools were 0.19 standard deviations above the district average. Because of that disparity, we limited the schools included in the control group to the lower half of all regular district schools. Those 71 schools had prior test scores that were 0.15 standard deviations below the district average, a level of performance much closer to those at the treated schools. Restricting the control group in this way allows us to make a cleaner, if not an exact, comparison while maintaining a sufficient number of schools to be able to detect sizeable management impacts at conventional levels of statistical significance.

Table A1 shows summary statistics for four groups of schools: for-profit, nonprofit, all district schools, and the low-performing regular district schools.[8] The half of regular district schools with the lowest pre-treatment test scores have average student characteristics that are more similar to the privately managed schools than the full set of regular district schools. However, some differences remain.

Method

The Philadelphia intervention does not provide the opportunity for a random assignment study of the impact of for-profit and nonprofit management. Schools assigned to intervention status were not chosen randomly but selected on the grounds that they were in greatest need of an intervention. We instead estimate the impact of attending a for-profit or nonprofit privately managed school (relative to that school had it remained under district management) using a quasi-experimental research design that employs "difference-in-differences" analysis.[9] The treatment groups consist of schools managed by each type of provider, and the control group includes the regular public schools with test scores below the median for all regular district schools, as discussed above. To identify the effect of treatment, we calculate in a single estimating equation, after adjusting for previous student test-score performance, student characteristics, and school fixed effects, the difference between average test scores at treated and control schools before and after the intervention began. So, for example, if test-score gains at the schools treated by for-profit management were 20 percent of a standard deviation higher than gains before treatment, while comparably measured gains at the control schools were only 15 percent of a standard deviation higher, the estimated effect of for-profit management would be the difference between them, or 5 percent of a standard deviation. Since the data allow for analysis over a longer period of time than just one year, we are able to identify the effect of attending a privately managed school over a six-year time span.

As just stated, the models employed in this analysis estimates gains in student test-score performance at the for-profit and nonprofit schools from one year to the next relative to what would have been expected had they remained under regular district management. The use of gain scores (or, in the case of Model I, test-score levels with test-score levels from the previous year included as covariates) thus provides an estimate of treatment effects based on the extent to which students at each school do better or worse than would be expected, given their initial test scores.

We estimate three different models. In the first, Model I, students' test scores in a given year and subject are the dependent variable; the previous year's test scores in both subjects are included as control variables. A limited set of student demographic characteristics (race, gender, special education status, and limited English proficiency status) are also included as covariates. However, it was not possible to control for parental education, parental income, student commitment or other educationally significant student background characteristics. Model II is identical to Model I except that, in Model II, the dependent variable is a test-score gain and thus prior-year test scores are not included as controls (because they are incorporated into the dependent variable). Model III is identical to Model II in all respects except that Model III includes student fixed effects instead of controls for the limited set of student background characteristics that were available. The inclusion of student fixed effects allows for the estimation of treatment effects by comparing each student's test-score gains to his or her own performance at other points in time. This allows the model to account for changes in the composition of the schools' student populations over time that cannot be accounted for by the limited set of student characteristics for which information was available. For this reason Model III is our preferred model.

All three models also control for students' peer characteristics (student characteristics aggregated to the school*grade*year level, since we cannot match students to their classrooms) and for whether students have recently moved from one school to another. Student movement among schools may be either structural—i.e., the student moves from an elementary to a middle school—or non-structural—e.g., the student changes school because the family moves to a different neighborhood, even while the student remains at an elementary school. Controls are included for these two different types of school changes.[10]

The equation for Model III, our preferred model because it accounts for changes in unobservable time-invariant student characteristics, is:

$$\begin{aligned}\Delta Y_{igst} &= \alpha + \beta_1 \mathrm{FP1}_{st} + \beta_2 \mathrm{FP2}_{st} + \beta_3 \mathrm{FP3}_{st} + \beta_4 \mathrm{FP4}_{st\,t} + \beta_5 \mathrm{FP5}_{st} + \beta_6 \mathrm{FP6}_{st} + \\ &\beta_7 \mathrm{NP1}_{st} + \beta_8 \mathrm{NP2}_{st} + \beta_9 \mathrm{NP3}_{st} + \beta_{10} \mathrm{NP4}_{st} + \beta_{11} \mathrm{NP5}_{st} + \beta_{12} \mathrm{NP6}_{st} + \\ &\lambda_1 \mathrm{SM}_{\mathrm{it}} + \lambda_2 \mathrm{NM}_{\mathrm{it}} + \delta_i + \eta_{gst} + \mu_s + \theta_{gt} + \varepsilon_{igst}\end{aligned}, \qquad (1)$$

where ΔY_{igst} is the decile-standardized test-score gain of student *i* in grade *g* in school *s* in year *t;* α is a constant; FP1 through FP6 are treatment dummy variables (with coefficients β_1 through β_6) indicating whether the school the student was attending in year *t* was in its first through sixth year of for-profit management (respectively); NP1 through NP6 are treatment dummy variables (with coefficients β_7 through β_{12}) indicating whether the school the student was attending in year *t* was in its first through sixth year of nonprofit management (respectively);[11] SM and NM (with coefficients λ_1 and λ_2) are dummy variables indicating whether the student has made a structural or non-structural move, respectively, from the previous year;[12] δ is a vector of student fixed effects; η is a vector of peer characteristics (student characteristics aggregated to the school*grade*year level); μ is a vector of school fixed effects; θ is a vector of grade-by-year fixed effects; and ε is the error term.[13]

As explained above, Model II replaces the student fixed effects in Model III with controls for student characteristics. Model I is the same as Model II, except that the dependent variable is the test-score level, and prior-year test scores (in both subjects, along with their squared and cubed terms) are included on the right-hand side of the equation. The three models are run separately for math and reading scores, and standard errors are adjusted for clustering by schools.[14]

We use a gain score as the dependent variable in Models IIII rather than controlling for the prior-year score on the right-hand side of the equation because controlling for an endogenous lagged dependent variable in a model that includes student fixed effects produces biased estimates (Angrist and Krueger 1999). One drawback of Models II and III is that they assume that the coefficient on the prior-year test score is equal to one when the true relationship between prior-year and current-year test scores may be otherwise. Of particular concern, students who received a low score on a test in a given year made, on average, larger gains the next year than did students who received a higher initial score. We address that issue in Models II and III by calculating decile-standardized gain scores that limit comparisons among students to just those whose initial scores fell within the same decile of the distribution (Hanushek et al. 2005). Decile-standardized gain scores are standardized to have a mean of zero and standard deviation of one within each grade, year, and decile of the prior-year test-score distribution.[15] We also present results that allow the coefficient on prior-year scores to vary (Model I). However, it is not possible to include student fixed effects in such a model (Angrist and Krueger 1999).

The inclusion of student fixed effects in Model III means that students are compared to themselves over time. Since estimates of treatment effects are based only on those students for whom at least two gain scores are observed, the student must have at least three test-score level observations. More than three observations are required to compute two gain scores if the three level scores are not consecutive. In our analysis, there are 68,677 students for whom we can compute at least two gain scores in math, of whom 20,696 have two gain scores, 20,011 have three gain scores, 14,331 have four gain scores, 13,088 have five gain scores, 548 have six gain scores, and 3 have seven gain scores. The 46,875 students for whom we can only compute one gain score are excluded from Model III, but included in Models I and II.[16]

Although thousands of student observations are available, those in the treatment condition are clustered within only 30 for-profit schools and 16 nonprofit ones. As a result, the annual impacts of the intervention must be as much as 20 to 30 percent of a standard deviation in order to be detected at conventional levels of statistical significance. That size of an annual impact is seldom detected for large-scale structural interventions in education. For that reason, we also discuss large impacts that fall short of statistical significance when the pattern of results is consistent over the six year time period.

Results

Our main results are presented in Table 2. For each model, test subject, treatment group (for-profit and nonprofit), and year after private management began, we present the estimated treatment effect along with its standard error. As mentioned previously, standard errors are robust to the clustering of students within schools. At the bottom of the table, we list the range of the number of student observations (separately for the treatment and control groups) contributing to each year of treatment effects.[17]

Impacts Relative to District Management

Nonprofit management

Math. Students appear to have learned less in mathematics in schools under nonprofit management than would have happened had the schools been under district management. During the six years of treatment, students at schools managed by the nonprofits were learning, on average, 21 percent of a standard deviation less (each year) in mathematics than they would have had their school remained under district management. However, the estimated impacts, though sizeable, fall short of statistical significance when estimated by our preferred model, Model III. Only in the first year of the intervention were the negative impacts statistically significant. Models I and II reveal large and statistically significant negative impacts. However, the models do not control for unobservable characteristics. They take into account students' initial test scores as well as their race, gender, eligibility for special education and limited English proficiency status, but the models may not capture other background characteristics—family education, family income, household formation, student commitment, and so forth—that could be critical to student achievement. Still, it is disconcerting to notice that the impacts from Model III, though not statistically significant, also estimate negative impacts that are, in the case of mathematics, as large as those obtained from Models I and II. In sum, the findings, though not definitive, are indicative of an adverse impact of nonprofit management on math performance.

Reading. The average annual impacts on reading performance were, on average, a negative 9 percent of a standard deviation. Only one of the estimations from Model III was statistically significant, however. Nonetheless, there is very little likelihood of a positive impact in either subject, as all estimations from all models in all but one year estimate negative impacts.

For-profit management.

Math. For-profit management seems to have had a positive impact on math performance. Our preferred Model, Model III, estimates large, positive impacts in the second year and all subsequent ones. On average for the six years, annual impacts were a positive 0.25 standard deviations. Only in the first year of the intervention were the positive impacts statistically insignificant. In subsequent years impacts range from 0.23 in year two to a high of 0.35 standard deviations in year five and an impact of 0.29 standard deviations in year six. When estimated by Models I and II, the size of the impacts on math performance remain positive in all years after the first year of the intervention. But they are considerably smaller and no longer are estimated at a statistically significant level. That may be due to the fact that neither of these models take into account changes in unobserved student characteristics, such as parental education, parental income and student commitment. However, the small impacts observed by Models I and II could also be due in part to the population for whom the impacts are estimated, a possibility considered further below.

Reading. In reading, all Model III estimates of student performance at the schools under for-profit management relative to what would have happened under district management are positive. On average for the six years, positive impacts are 10 percent of a standard deviation. After the first year following the intervention, estimates range between 0.09 and 0.15 standard deviations. However, none of the estimations are statistically significant. Impacts estimated by Models I and II are smaller, less consistently positive, and always fall short of being detected at conventional levels of statistical significance. Their smaller size may, once again, be due to the fact that the models are unable to control for unobserved student characteristics or to the fact that a different student population was observed.

<u>Differential Impacts of Nonprofits vs. For-profits</u>

Math. The clearest finding—one that is consistent across all three models—is that for-profit management was more effective at raising student achievement in math than nonprofit management. As estimated by Model III, differential impacts for the six years were, on average, 0.46 standard deviations. The impacts ranged between 0.40 and 0.36 standard deviations in years one and two to 0.53 and 0.54 standard deviations in years five and six. In all six years estimations are statistically significant.[18] Differential impacts remain statistically significant when estimated from Models I and II, though the size of the estimated impact is smaller. As estimated by Models I and II, average differential impacts over the six years were 0.17 and 0.24 standard deviations, respectively. Robust to all specifications, the for-profits appear to have out-performed the nonprofits in the instruction of mathematics.

Reading. For-profits also outperformed the nonprofits in securing gains in reading achievement. According to Model III, the differential impacts of the two sectors averaged 0.19 standard deviations for each of the six years, a difference that was statistically significant at better than the .1 level in all but one year. Smaller but still statistically significant differentials were estimated by Models I and II in all years but one.

Informally Adjusted Estimations

All estimations are based on a subset of students at schools in the Philadelphia school district. No information is available for students in kindergarten and first grade. Only a small amount of test-score information is available for students in second grade, and for many other students only one year of test score information is available in a given year. So none of the estimations provided in this study can be extrapolated beyond the specific group of students for whom the necessary test-score information is available. As mentioned above, Model III estimations require three years of test-score observations. Only then can one estimate impacts by taking into account a students' unobserved fixed characteristics. Fortunately, we have that information for more than 68,000 observations. But for nearly another 47,000 observations, we could observe impacts using only the less desirable Models I and II that required only two years student test-score performance. Even these models fall far short of providing an estimate of the performance of all the students at the for-profit, nonprofit and district schools.

This raises the possibility that the students included in the estimations are not representative of all the students participating in the intervention. Perhaps the large impacts observed by Model III as compared to Models I and II are due not to an improved estimation but to the smaller set of students included in the estimation. Table A5 shows the average characteristics of students in each model, separately for 2002 and 2008 (but combining the treatment and control groups). In 2008, when the data are more complete, the group of students included in Model III are very similar, on average, to students in Models I and II and all students. In 2002, prior-year test scores are available for fewer students because fewer grades were tested in 2001. As a result, Model III students have noticeably lower test scores than Model I and II students. Table A6 shows that this is particularly pronounced at the for-profit schools. At the low-performing regular district schools, Model III students have math scores than are 0.05 lower than Model I and II students. At the for-profit schools, the same difference is 0.10. Thus, restricting the sample to the Model III students bases estimations on students that have pre-treatment test scores that are an additional 0.05 lower than the control group.

To see whether changes in the composition of the students included in the three estimations could explain the larger Model III impacts, a restricted version of Model II was used to estimate impacts for only those students who could be included in the Model III analyses. Table A2 of the Appendix displays results from the unrestricted Model II reported in the main text, results from Model II when restricted to just those students included in the Model III estimate, and results from Model III.

It is possible for the findings from this exercise to take one of two extreme patterns. At one extreme, the restricted estimates from Model II could be identical to Model II unrestricted estimates. If so, one can infer that the students included in Model III are a representative sample of all observations used to estimate unrestricted Model II. At the other extreme, Model II restricted estimates might be identical to Model III estimates. If so, one may infer that Model II observables capture all the unobservables that the fixed effects analysis is designed to detect. Unrestricted Model II would then be preferred over Model III, because it is based on a larger set of observations.

If results from the three estimations fall between these two extremes, one might infer that any difference between the restricted and the unrestricted Model II models can be subtracted from the Model III estimate to obtain an approximation of the Model III impacts on the larger set of students included in Model II. Although this requires making a number of restrictive assumptions about the nature of the underlying data, we shall use this analysis in order to present informally adjusted estimations that indicate the likely magnitude of the impacts for as many observations as possible.

In the discussion below, we refer to such results as adjusted impacts. However, it must be emphasized that these adjustments are informal. The best available estimations are Model III estimates for those students who had three years worth of test scores. The extent to which one can generalize from those results to all students participating in the intervention remains uncertain, however.

As can be seen in Table A2, adjusted results sometimes differ noticeably from those obtained from Model III, though they do not alter the overall findings that for-profit management proved superior to district management and, even more clearly, to nonprofit management. Math results for nonprofits vary little between Model II and III, which implies that unobservable characteristics were well captured by the observable ones, and the population included in Model III was representative of the larger population used to estimate Model II. But for reading, results for the restricted sample often fall approximately halfway between results from Model II unrestricted and Model III. When adjusted, impacts on reading performance of nonprofit management is just -0.09 standard deviations.

The size of the positive impacts also attenuate when uses the same methodology to generalize for-profit impacts to the larger student population included in Model II estimates. In reading, average annual impacts for the six years are 0.04 standard deviations, while in math they are 0.12 standard deviations.

Size of the Impacts

One way of obtaining a rough sense of the size of the impacts being reported is to calculate them in terms of years of schooling. In Philadelphia, one standard deviation in test-score gains on the math and reading components of the Terra Nova (the test taken almost exclusively during three of the six years of the post-treatment period) is equivalent to 2.4 and 3.6 years of learning, respectively, from grades 2 through 8.[19] That implies that 42 percent of a standard deviation in math gains is equivalent to one year of learning within the Philadelphia school district. In reading, 28 percent of a standard deviation in gains is equivalent to one year of learning. All calculations are rough, however, as the Terra Nova was only one of three tests that were administered.

Averaging across each of the estimations from Model III for the six post-treatment years, the negative nonprofit effect of 21 percent of a standard deviation in math indicates roughly that students learned 50 percent of a year's worth of learning *less* each year. The average annual negative effect of 9 percent of a standard deviation in reading roughly indicates 32 percent of a year's worth of learning *less* each year. In other words, the impacts are sizeable, even though the estimations usually fall short of conventional levels of statistical significance. Adjusted results in reading are even more severe. The negative impact on learning increases from 32 percent to 50 percent of a year's worth of learning. In math, the loss in learning is 43 percent of a year's worth when results are adjusted.

The average annual for-profit positive effect in math of 25 percent of a standard deviation is approximately equivalent to an *additional* 60 percent of a year's worth of learning. These are thus large, statistically significant, educationally meaningful impacts. When the adjustment is made, average impacts are 29 percent of a year's worth of learning each year, still a meaningful impact.

In reading, average annual impacts average 36 percent of a year's worth of learning. When adjusted, they are 14 percent of a year's worth of learning. The reading impacts fall short of statistical significance, however.

Finally, the differential impact of for-profit and nonprofit management is very large. Using the same average estimates as above, students in schools under for-profit management were learning the equivalent of well over a year's worth of learning in math and 70 percent of a year's worth of learning in reading *each year* than students in schools under nonprofit management. The math differences are always statistically significant. In reading, the differences between for-profit and nonprofit effects are statistically significant in five out of six years.[20] When adjusted, average annual differences are 72 percent of year's worth of learning in math and 64 percent of a year's worth of learning in reading.

2008 Decision to Re-assume Control of Five For-Profit Schools

Our analysis provides compelling evidence that schools do much better under for-profit than nonprofit management. Year after year, students were learning more in reading and math, apparently by very wide margins if they attend a school under for-profit rather than nonprofit management. Yet the district reassumed control of only one school under nonprofit management while not renewing the contract of five for-profits. To ascertain whether that decision had a strong educational basis in the district's own test score data base, we used the same methodology to estimate the impact on student learning of the five schools for whom the for-profit management contract had been terminated.

The Model III results, as reported in Table A4, are mixed but provide little support for the district's decisions. On the one hand, the reading performances of students at the five schools under for profit management was, on average, 18 percent of a standard deviation below what could have been expected had the schools been under district management, a difference that is statistically significant in three of the six years. Also, nonprofit schools whose contract was not revoked had an impact on reading performance that was 10 percent of a standard deviation more positive than that of the for-profit schools whose contract was terminated. However, that difference is not statistically significant in any year.

On the other hand, the math performance of students at the for-profit schools was 35 percent of a standard deviation higher than would have been the case had the schools been under district management. Also, it was 56 percent of a standard deviation higher than would have been the case had the schools been under nonprofit management. Those large differences are statistically significant in most years.

Model III estimates indicate that nonprofits out-performed the five for-profits in reading in five out of six years, although the difference was only 3 percent of a standard deviation in 2008 and in no year were the differences statistically significant. In math, the five for-profits had strongly positive impacts in all years, while the nonprofits had decidedly negative ones, leading to very large, statistically significant differences between the two groups of schools in all years. The large differences in math clearly offset the statistically insignificant reading differences.

If the Philadelphia school district cares only about reading results, and places no weight on math ones, our data could be used to support the policy choice that was made, provided no attention is paid to the statistical insignificance of the reading finding. But if the two subjects are given equal weight in evaluating a school, our results provide no support for the decisions taken by the school district with respect to for-profit and nonprofit management.[21]

Discussion

Care should be taken before generalizing from the Philadelphia experience concerning the relative advantage of for-profit and nonprofit management. It is possible that for-profit entities have a greater vested interest in enhancing student achievement, because only in that way are they likely to survive over the longer run. But other factors in Philadelphia could easily account for the same result. The two main for-profit providers had much more experience with school management than did any of the nonprofit organizations. The nonprofits seem to have been selected as much for their strong political base as for any history of effectiveness at delivering educational services. Others have reported that newly formed charter schools under both for-profit and nonprofit management appear to become more effective as they gain in experience (Gill et al. 2001). That could easily account for the pattern of results reported here. Still, it is disconcerting to discover that impacts of non-profits compared unfavorably with those of the for-profits six years after the intervention began, presumably a long enough period for an entity to learn from experience.

Appendix: Propensity Score Analysis

To ascertain whether our estimates of for-profit and nonprofit management impacts are biased by the remaining differences in average school characteristics between the treatment and control groups (see Table 1), we also estimated treatment effects using a smaller control group selected using propensity scores.

First, using a probit model, we used school-level average student characteristics in 2001 and 2002 (special education status, race, and test scores) to predict whether the school was a for-profit or regular public school (the nonprofits, restructured, and sweet sixteen schools were excluded). The propensity scores are the predicted probabilities from this probit model. We then selected as the control group the 30 regular public schools with the highest propensity scores. As Table A1 shows, the smaller control group is more similar to the treated schools than the control group used in our main results (regular public schools in the bottom half of the pre-treatment test-score distribution). We then repeated this procedure with the nonprofits (excluding the for-profits) to select a control group of 30 regular district schools to be compared to the nonprofits. Note that, in the propensity score analysis, the control group used to estimate the for-profit effects is different from the control group used to estimate the nonprofit effects.

We then ran our three models, separately for the for-profits and nonprofits (unlike in the main results, where for-profit and nonprofit results are estimated simultaneously).

The results from the propensity score analysis are reported in Table A3. Estimations are so imprecise that standard errors become very large. The standard errors from Model III estimations vary between 0.12 and 0.24, precluding the identification of impacts ranging between a quarter to a half a standard deviation. No conclusions can be drawn from the analysis other than that the general pattern of impacts resemble those reported in the main analyses. In general, nonprofits appear to have negative impacts on math performance. while for-profits appear to have positive ones. Reading impacts hover near zero. Nothing reaches a level of statistical significance, however.

References

Angrist, Joshua D. and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics," in the *Handbook of Labor Economics*, Vol. 3A, Orley Ashenfelter and David Card, eds., Elsevier, pp. 1278-1366.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119(1), pp. 249-75.

Brandl, John E. 1998. *Money and Good Intentions are Not Enough: Or Why a Liberal Democrat Thinks States Need Both Competition and Community.* Washington: Brookings Institution Press.

Brandl, John E. 2006. "Choice, Religion, Community and Educational Quality." In Robert C. Enlow and Lenore T. Ealy. *Liberty and Learning: Milton Friedman's Voucher Idea at Fifty*. Washington: CATO. pp. 25-34.

Christman, Jolley Bruce, Eva Gold, and Benjamin B. Herold. 2006. "Privatization 'Philly Style': What Can Be Learned from Philadelphia's Diverse Provider Model of School Management." *Research Brief.* Philadelphia, Pa.: Research for Action, (June). Accessed March 27, 2007: http://www.researchforaction.org/publication/details/37.

Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Coulson, Andrew J. 1999. *Market Education: The Unknown History.* New Brunswick, N. J.: Transaction.

Dean, Mensah M. 2008. "Private Managers Lose 6 Schools," *Philadelphia Daily News*, June 19.

Dynarski, Susan M. 2003. "Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion." *The American Economic Review*, 93(1), pp. 279-288.

Friedman, Milton. 1955. "The Role of Government in Education" in Robert Solow, ed. *Economics and the Public Interest.* New Brunswick, New Jersey: Rutgers University Press.

"Gamble and Huff." 2008. Wikipedia entry, available http://en.wikipedia.org/wiki/Gamble_and_Huff, accessed September 7, 2008.

Gill, Brian P., P. Michael Timpane, Karen E. Ross, and Dominic J. Brewer. 2001. *Rhetoric versus Reality: What We Know and What We Need to Know About Vouchers and Charter Schools.* Pittsburgh, Pennsylvania: RAND Education.

Gill, Brian, Ron Zimmer, Jolley Christman, and Suzanne Blanc. 2007. "State Takeover, School Restructuring, Private Management, and Student Achievement in Philadelphia," RAND Corporation and Research for Action.

Hanushek, Eric, John Kain, Daniel O'Brien, and Steven Rivkin. 2005. "The Market for Teacher Quality." NBER Working Paper No. 11154.

Hanushek, Eric A. and Ludger Wößmann. 2006. "Does Educational Tracking Affect Performance and Inequality? Differences- in-Differences Evidence Across Countries." *The Economic Journal* 116(510), pp. C63-C76.

Hassel, Brian. 2003. "The Future of Charter Schools." In Paul E. Peterson, ed. *The Future of School Choice*. Hoover Institution Press.

Herold, Benjamin B., and Morgan Riffer. 2005. "The 'Original 86': Tracking Changes in the 'Original 86' Philadelphia Public Schools Initially Targeted for Intervention Following the 2001 State Takeover," Philadelphia, Pa.: Research for Action. Accessed October 7, 2007: http://www.researchforaction.org/publication/details/152.

Hill, Cynthia D. and David M. Welsch. 2008. "For-profit versus Not-for-profit Charter Schools: An Examination of Michigan Test Scores." *Education Economics*.

Hochschild, Jennifer and Nathan Scovronick. 2003. *The American Dream and the Public Schools.* Oxford University Press.

Levin, Henry. 2001. "Bear Market: The Recent Entry of For-profit Firms into the K-12 Arena is an Intriguing Trend." *Education Next* (Spring 2001), pp. 6-15.

Matthews, Jay. 2003. "The Philadelphia Experiment," Education Next (Winter), pp. 50-56.

Mead, Sara. 2007. "Easy Way Out: Restructuring Usually Means Little has Changed," *Education Next* (Spring), pp. 52-56.

Meyer, Bruce D. 1995. "Natural and quasi-experiments in economics." *Journal of Business and Economic Statistics* 13(2), pp. 151-161.

Merrifield, John. 2001. *The School Choice Wars.* Lanham, MD: Scarecrow.

Peterson, Paul E. 2007. "School Reform in Philadelphia: A Comparison of Student Achievement at Privately-Managed Schools with Student Achievement in Other District Schools," Program on Education Policy and Governance Paper 07-03.

Peterson, Paul E. and Matthew M. Chingos. 2007. "Impact of For-Profit and Non-Profit Management on Student Achievement: The Philadelphia Experiment," Program on Education Policy and Governance Paper 07-07.

School District of Philadelphia. 2004. “School District of Philadelphia Outperforms State Averages for PSSA Progress.” Press Release. (July 22).

Endnotes

[1] We wish to thank the School District of Philadelphia's Office of Accountability, Assessment, and Intervention for providing the information on student performance. We appreciate the comments on earlier drafts of this paper by Terry Moe, Eric Hanushek, Steve Rivkin, Martin West and the participants in the conference on *Economic Incentives: Do They Work in Education?* sponsored by Harvard University's Program on Education Policy and Governance and CES-ifo Economic Research Institute held in Munich, Germany, May 16-17, 2008. Antonio Wendland and Ron Berry provided staff assistance. The research for this paper is supported in part by the National Research and Development Center on School Choice, Competition, and Achievement, which is funded by the Department of Education's Institute of Education Sciences (R305A040043) as well as by grants from the John M. Olin Foundation, the Lynde and Harry Bradley Foundation, and Edison Schools. The findings and interpretation are the sole responsibility of the authors.

[2] Another 21 schools were assigned to be restructured by a newly created Office of Restructured Schools (ORS), a special office within the school district itself (Herold and Riffer 2005). A high school was contracted out to a for-profit provider in 2004.

[3] Hill and Welsch (2008) use school-level data to compare for-profit and nonprofit charter schools in Michigan. They find no difference in test scores between the two management types.

[4] Similar results are obtained when Chancellor Beacon schools are excluded from the analysis.

[5] The study is limited to schools that began treatment in the fall of 2002. Two schools were contracted to Edison Schools in 2005, and another school was contracted to Victory Schools in the fall of 2003. To ensure direct comparisons of the same cohorts, we exclude those schools from our analysis. Additionally, the data set includes school enrollment information on non-tested grades (K-12), which allowed us to determine whether a student was new to his or her school during a given year.

[6] Throughout this paper we refer to school years by the calendar year in which spring falls, as that is when tests are administered. Thus, 2001 refers to school year 2000-2001, and so forth.

[7] At the time schools were given to private managers, other schools in Philadelphia, known as restructured schools, were reorganized by the school district itself, and still other schools, known as the "sweet sixteen," were given additional resources. We include these schools in our analysis to detect impacts of the district's own restructuring initiative, but they do not form part of the control group. We found no statistically significant impacts of these interventions.

[8] Control schools were selected on the basis of all students' test scores in the two pre-treatment years (2001 and 2002). Table A1 also shows summary statistics for control schools selected for the propensity score analysis, which we discuss later in this paper.

[9] See Cook and Campbell (1979, pp. 214-18). For another general reference on difference-in-differences estimation, see Meyer (1995). Two recent papers in the education literature that use this method are Dynarski (2003) and Hanushek and Wößmann (2006).

[10] The performance of students who leave the Philadelphia school district cannot be tracked. Also, students who are in their first year in the Philadelphia schools cannot be included in the analysis because the prior-year scores necessary for calculating gain scores are not available.

[11] We also include sets of dummy variables identifying restructured and sweet sixteen schools in their first through fourth years of treatment.

[12] A student is determined to have made a structural move if he or she is in a different school from the previous year and his or her grade this year exceeds the maximum grade of the school attended during the previous year. Student are determined to have made a non-structural move if they are currently in a different school from the one attended the previous year and their grade level in the current year does not exceed the maximum grade level of the school attended during the previous year.

[13] The model is equivalent to one with a difference-in-differences interaction term (a dummy variable indicating whether the observation is a treated unit in the period after treatment has begun) for each treatment type (for-profit and nonprofit) with this term interacted with the year dummies from the post-treatment period (except for the year dummy from the first post-treatment year). The advantage of the model used here is that it allows one to directly estimate treatment effects (and standard errors) for each year post-treatment, while the standard model with interaction terms would require the addition of the main effect to the coefficient on each of the interaction terms in order to obtain an estimate of treatment effects for that year.

[14] Bertrand et al. (2004) show that the difference-in-differences method is unlikely to produce biased standard errors as long as the number of units (in our case, schools) is more than ten.

[15] When we use regular gain scores instead of decile-standardized gain scores we obtain qualitatively similar results.

[16] The observations for reading scores are similar in number.

[17] See Appendix for a supplementary propensity score analysis. The estimated impacts of the restructured and sweet sixteen schools (not shown) are generally small and always statistically insignificant in both subjects in all years.

[18] We test the statistical significance of differences in treatment effects between for-profit and nonprofit managers using an F test of the null hypothesis of no difference between the for-profit and nonprofit treatment coefficients in a given year.

[19] We use scale scores (i.e., not standardized) to make these calculations, as the Terra Nova scale scores can be compared over time (unlike standardized scores, which by definition have a mean of zero). For each pair of adjacent grades, we calculate the average difference in scale score

between the years, which is the test-score gain of the average student in the Philadelphia school district who stayed on grade level (i.e., was not held back a year). We then take the standard deviation of the gain scores (recall that our results are all calculated as percentages of standard deviations of gain scores) and divide it by the mean gain score. The resulting number is the number of years of learning that is equivalent to a one-standard-deviation change in the gain score. We then take the simple average, separately for math and reading, of this calculation for each pair of grades between 2 and 8.

[20] A prior investigation conducted by RAND-Research for Action (RAND-RFA) reported no impact of private management on student test-score performance in either reading or mathematics (RAND-RFA, 2006). Our study differs from that one in a number of technical ways. Most importantly, the RAND-RFA study combines the two types of private management, while we estimate separately the impact of for-profit and nonprofit managers, finding sharp contrasts between them.

[21] Results from Models I and II also do not support the school's decisions. In math, the five for-profits outperformed the nonprofits every year. In reading, the for-profits performed better in years 1 and 2, worse in years 4 and 5, and about the same in years 3 and 6.

Table 1. Tests Administered by Grade and Year Tested (in Spring)

Grade	2001	2002	2003	2004	2005	2006	2007	2008
2			Terra Nova	Terra Nova	Terra Nova	Terra Nova	Terra Nova	Terra Nova
3	Stanford 9	Stanford 9	Terra Nova	Terra Nova	Terra Nova	PSSA	Terra Nova & PSSA	PSSA
4	Stanford 9	Stanford 9	Terra Nova	Terra Nova	Terra Nova	PSSA	Terra Nova & PSSA	PSSA
5	PSSA	PSSA	Terra Nova & PSSA	Terra Nova & PSSA	Terra Nova & PSSA	PSSA	Terra Nova & PSSA	PSSA
6			Terra Nova	Terra Nova	Terra Nova	PSSA	Terra Nova & PSSA	PSSA
7	Stanford 9	Stanford 9	Terra Nova	Terra Nova	Terra Nova	PSSA	Terra Nova & PSSA	PSSA
8	PSSA	PSSA	Terra Nova & PSSA	Terra Nova & PSSA	Terra Nova & PSSA	PSSA	Terra Nova & PSSA	PSSA

Notes: The Stanford 9 and Terra Nova are nationally normed tests. The Pennsylvania State System of Assessment (PSSA) is the primary vehicle for holding schools accountable for improving student learning in Philadelphia.

Table 2. Effects of Private Management on Test Scores

		Math			Reading		
		Model I	**Model II**	**Model III**	**Model I**	**Model II**	**Model III**
Nonprofit	One-year	-0.17 [0.06]***	-0.25 [0.08]***	-0.29 [0.16]*	-0.15 [0.04]***	-0.24 [0.06]***	-0.26 [0.11]**
	Two-year	-0.12 [0.05]**	-0.16 [0.08]**	-0.13 [0.14]	-0.11 [0.05]**	-0.15 [0.07]**	-0.04 [0.10]
	Three-year	-0.13 [0.06]**	-0.18 [0.09]**	-0.18 [0.15]	-0.15 [0.05]***	-0.21 [0.07]***	-0.11 [0.12]
	Four-year	-0.16 [0.07]**	-0.22 [0.10]**	-0.20 [0.16]	-0.11 [0.06]*	-0.11 [0.09]	0.01 [0.13]
	Five-year	-0.13 [0.07]*	-0.17 [0.10]*	-0.18 [0.16]	-0.12 [0.06]**	-0.19 [0.08]**	-0.10 [0.13]
	Six-year	-0.16 [0.07]**	-0.25 [0.11]**	-0.25 [0.18]	-0.11 [0.07]*	-0.14 [0.09]	-0.05 [0.11]
For-Profit	One-year	-0.04† [0.06]	-0.08† [0.09]	0.11††† [0.12]	-0.03†† [0.05]	-0.06†† [0.07]	0.01†† [0.08]
	Two-year	0.03††† [0.04]	0.05††† [0.06]	0.23††† [0.12]*	0.01†† [0.03]	0.05††† [0.05]	0.13† [0.08]
	Three-year	0.02††† [0.04]	0.04†† [0.05]	0.25††† [0.10]**	-0.01††† [0.04]	0.00††† [0.06]	0.09† [0.08]
	Four-year	0.02†† [0.05]	0.05†† [0.07]	0.28††† [0.13]**	-0.02 [0.04]	0.02 [0.06]	0.10 [0.08]
	Five-year	0.05††† [0.05]	0.10††† [0.08]	0.35††† [0.13]***	0.00†† [0.05]	0.00†† [0.07]	0.11† [0.09]
	Six-year	0.05††† [0.05]	0.07††† [0.08]	0.29††† [0.13]**	0.02†† [0.04]	0.06†† [0.07]	0.15† [0.10]
Time-invariant controls?		Yes	Yes	No	Yes	Yes	No
Time-varying controls?		Yes	Yes	Yes	Yes	Yes	Yes
Student Fixed Effects?		No	No	Yes	No	No	Yes
Nonprofit Schools		16	16	16	16	16	16
For-Profit Schools		30	30	30	30	30	30
Control Schools		71	71	71	71	71	71
Nonprofit Students		1,915–4,797	1,933–4,810	812–3,985	1,929–4,763	1,985–4,788	832–3,954
For-Profit Students		5,318–12,391	5,379–12,802	2,404–11,425	5,328–12,232	5,466–12,304	2,409–11,118
Control Students		9,843–24,194	9,912–24,535	5,196–21,860	9,851–23,927	9,993–24,002	5,229–21,595

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%; ††† (††) [†] indicates that the for-profit effect is significantly different from the nonprofit effect at 1% (5%) [10%]; robust standard errors adjusted for clustering within schools appear in brackets. Results for math and reading are estimated separately. Untreated schools include those in the bottom half of the (school aggregate) pre-treatment test-score distribution. Time-invariant controls include students' race and gender. Time-varying controls include students' special education and limited English proficiency status as well as indicator variables for whether the student made a structural or nonstructural move from the previous school year. In Model I, the dependent variable is the test-score level and prior-year test scores are included as controls. In Models II and III, the dependent variable is the decile-standarized test-score gain. All models include school fixed effects, grade-by-year fixed effects, and students' peer characteristics.

Table A1. Summary Statistics of Student Characteristics Prior to Treatment

	For-Profit Schools	Nonprofit Schools	Regular District Schools (All)	Regular District Schools (Lower Performing)	Propensity Score Control Group (For-Profit)	Propensity Score Control Group (Nonprofit)
Number of Schools	30	16	141	70	30	30
% Special Ed 2001	13.7%	12.3%	21.3%	17.3%	15.9%	13.5%
% Special Ed 2002	12.0%	12.1%	16.6%	13.5%	12.6%	11.3%
% Black 2001	80.1%	94.5%	57.0%	68.3%	80.8%	93.2%
% Black 2002	79.9%	95.1%	57.9%	69.2%	81.3%	94.7%
% Hispanic 2001	16.3%	2.7%	12.7%	12.7%	12.4%	1.9%
% Hispanic 2002	16.5%	2.7%	13.4%	13.5%	12.5%	1.6%
Math Score 2001	-0.41	-0.31	0.20	-0.15	-0.32	-0.26
Math Score 2002	-0.40	-0.24	0.21	-0.15	-0.30	-0.17
Reading Score 2001	-0.36	-0.28	0.17	-0.16	-0.30	-0.21
Reading Score 2002	-0.37	-0.22	0.18	-0.15	-0.26	-0.12

Notes: Math and reading scores are standardized by subject, test, grade, and year to have a mean of zero and standard deviation of one. The one school (regular district) with missing data in one of the pre-treatment years is excluded from this table.

Table A2. Student Fixed Effects vs. Restricted Sample

		Math			Reading		
		Model II Unrestricted	Model II Restricted	Model III	Model II Unrestricted	Model II Restricted	Model III
Nonprofit	One-year	-0.25 [0.08]***	-0.28 [0.13]**	-0.29 [0.16]*	-0.24 [0.06]***	-0.24 [0.09]***	-0.26 [0.11]**
	Two-year	-0.16 [0.08]**	-0.18 [0.12]	-0.13 [0.14]	-0.15 [0.07]**	-0.09 [0.09]	-0.04 [0.10]
	Three-year	-0.18 [0.09]**	-0.21 [0.12]*	-0.18 [0.15]	-0.21 [0.07]***	-0.15 [0.10]	-0.11 [0.12]
	Four-year	-0.22 [0.10]**	-0.24 [0.14]*	-0.20 [0.16]	-0.11 [0.09]	-0.05 [0.12]	0.01 [0.13]
	Five-year	-0.17 [0.10]*	-0.19 [0.13]	-0.18 [0.16]	-0.19 [0.08]**	-0.14 [0.11]	-0.10 [0.13]
	Six-year	-0.25 [0.11]**	-0.27 [0.16]*	-0.25 [0.18]	-0.14 [0.09]	-0.08 [0.11]	-0.05 [0.11]
For-Profit	One-year	-0.08† [0.09]	0.09††† [0.10]	0.11††† [0.12]	-0.06†† [0.07]	0.01†† [0.07]	0.01†† [0.08]
	Two-year	0.05††† [0.06]	0.16††† [0.09]*	0.23††† [0.12]*	0.05††† [0.05]	0.09†† [0.06]	0.13† [0.08]
	Three-year	0.04†† [0.05]	0.18††† [0.08]**	0.25††† [0.10]**	0.00††† [0.06]	0.06†† [0.07]	0.09† [0.08]
	Four-year	0.05†† [0.07]	0.18††† [0.10]*	0.28††† [0.13]**	0.02 [0.06]	0.07 [0.07]	0.10 [0.08]
	Five-year	0.10††† [0.08]	0.23††† [0.10]**	0.35††† [0.13]***	0.00†† [0.07]	0.06† [0.08]	0.11† [0.09]
	Six-year	0.07††† [0.08]	0.19††† [0.10]*	0.29††† [0.13]**	0.06†† [0.07]	0.11 [0.08]	0.15† [0.10]
Time-invariant controls?		Yes	Yes	No	Yes	Yes	No
Time-varying controls?		Yes	Yes	Yes	Yes	Yes	Yes
Student Fixed Effects?		No	No	Yes	No	No	Yes
Nonprofit Schools		16	16	16	16	16	16
For-Profit Schools		30	30	30	30	30	30
Control Schools		71	71	71	71	71	71
Nonprofit Students		1,933–4,810	812–3,985	812–3,985	1,985–4,788	832–3,954	832–3,954
For-Profit Students		5,379–12,802	2,404–11,425	2,404–11,425	5,466–12,304	2,409–11,118	2,409–11,118
Control Students		9,912–24,535	5,196–21,860	5,196–21,860	9,993–24,002	5,229–21,595	5,229–21,595

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%; ††† (††) [†] indicates that the for-profit effect is significantly different from the nonprofit effect at 1% (5%) [10%]; robust standard errors adjusted for clustering within schools appear in brackets. Results for math and reading are estimated separately. The middle column presents results using Model II but restricting the sample to include only the students that contribute to the Model III estimates. Untreated schools include those in the bottom half of the (school aggregate) pre-treatment test-score distribution. Time-invariant controls include students' race and gender. Time-varying controls include students' special education and limited English proficiency status as well as indicator variables for whether the student made a structural or nonstructural move from the previous school year. The dependent variable in both Models II and III is the decile-standarized test-score gain. All models include school fixed effects, grade-by-year fixed effects, and students' peer characteristics.

Table A3. Effects of Private Management on Test Scores, Propensity Score Analysis

		Math			Reading		
		Model I	**Model II**	**Model III**	**Model I**	**Model II**	**Model III**
Nonprofit	One-year	-0.11	-0.15	-0.09	-0.05	-0.09	-0.06
		[0.08]	[0.12]	[0.24]	[0.06]	[0.09]	[0.16]
	Two-year	-0.05	-0.07	0.02	-0.03	-0.03	0.13
		[0.06]	[0.09]	[0.20]	[0.06]	[0.08]	[0.13]
	Three-year	-0.08	-0.12	-0.03	-0.08	-0.10	0.05
		[0.07]	[0.11]	[0.22]	[0.05]	[0.08]	[0.16]
	Four-year	-0.13	-0.20	-0.12	-0.04	-0.02	0.14
		[0.07]*	[0.11]*	[0.22]	[0.06]	[0.09]	[0.17]
	Five-year	-0.09	-0.13	-0.07	-0.06	-0.11	0.02
		[0.07]	[0.11]	[0.23]	[0.06]	[0.09]	[0.16]
	Six-year	-0.14	-0.23	-0.18	-0.04	-0.04	0.13
		[0.08]*	[0.13]*	[0.25]	[0.07]	[0.10]	[0.15]
For-Profit	One-year	-0.09	-0.15	-0.02	-0.03	-0.05	-0.02
		[0.07]	[0.10]	[0.14]	[0.05]	[0.07]	[0.10]
	Two-year	0.00	0.01	0.13	0.02	0.05	0.07
		[0.05]	[0.07]	[0.16]	[0.04]	[0.06]	[0.12]
	Three-year	-0.03	-0.04	0.11	-0.01	0.00	0.02
		[0.04]	[0.07]	[0.12]	[0.05]	[0.07]	[0.11]
	Four-year	-0.03	-0.03	0.13	-0.01	0.00	0.02
		[0.06]	[0.09]	[0.16]	[0.06]	[0.08]	[0.13]
	Five-year	-0.02	-0.02	0.17	0.01	0.01	0.04
		[0.06]	[0.10]	[0.16]	[0.06]	[0.08]	[0.13]
	Six-year	-0.03	-0.07	0.08	0.02	0.05	0.08
		[0.06]	[0.09]	[0.15]	[0.05]	[0.08]	[0.12]
Time-invariant controls?		Yes	Yes	No	Yes	Yes	No
Time-varying controls?		Yes	Yes	Yes	Yes	Yes	Yes
Student Fixed Effects?		No	No	Yes	No	No	Yes
Nonprofit Schools		16	16	16	16	16	16
For-Profit Schools		30	30	30	30	30	30
Control Schools		30	30	30	30	30	30
Nonprofit Students		1,915–4,797	1,933–4,810	622–3,820	1,929–4,763	1,985–4,788	639–3,795
For-Profit Students		5,318–12,391	5,379–12,802	2,213–10,971	5,328–12,232	5,466–12,304	2,214–10,658
Control Student (Nonprofit)		4,107–10,011	4,136–10,078	1,332–8,114	4,112–9,917	4,187–9,952	1,346–8,073
Control Students (For-profit)		4,367–9,427	4,405–9,667	1,738–8,116	4,366–9,308	4,445–9,334	1,741–7,959

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%; robust standard errors adjusted for clustering within schools appear in brackets. Results for math and reading are estimated separately. Results for for-profit and nonprofit schools are estimated separately using different control groups (selected using propensity scores). Time-invariant controls include students' race and gender. Time-varying controls include students' special education and limited English proficiency status as well as indicator variables for whether the student made a structural or nonstructural move from the previous school year. In Model I, the dependent variable is the test-score level and prior-year test scores are included as controls. In Models II and III, the dependent variable is the decile-standarized test-score gain. All models include school fixed effects, grade-by-year fixed effects, and students' peer characteristics.

Table A4. Effects of Non-revoked Nonprofits and Revoked For-Profits on Test Scores

		Math Model I	Math Model II	Math Model III	Reading Model I	Reading Model II	Reading Model III
Nonprofits whose contracts were not revoked	One-year	-0.18 [0.06]***	-0.27 [0.09]***	-0.31 [0.16]*	-0.16 [0.04]***	-0.25 [0.06]***	-0.28 [0.11]**
	Two-year	-0.12 [0.05]**	-0.17 [0.08]**	-0.15 [0.15]	-0.11 [0.05]**	-0.14 [0.07]*	-0.04 [0.10]
	Three-year	-0.13 [0.06]**	-0.17 [0.09]**	-0.18 [0.15]	-0.15 [0.05]***	-0.20 [0.07]***	-0.10 [0.12]
	Four-year	-0.15 [0.07]**	-0.21 [0.10]**	-0.20 [0.16]	-0.10 [0.06]	-0.10 [0.09]	0.02 [0.14]
	Five-year	-0.11 [0.07]*	-0.15 [0.10]	-0.17 [0.16]	-0.11 [0.06]*	-0.17 [0.08]**	-0.09 [0.13]
	Six-year	-0.16 [0.07]**	-0.24 [0.12]**	-0.26 [0.18]	-0.10 [0.07]	-0.12 [0.09]	-0.05 [0.11]
For-Profits whose contracts were revoked	One-year	-0.12 [0.07]*	-0.18 [0.11]*	0.21††† [0.14]	-0.11 [0.06]*	-0.14 [0.07]**	-0.25 [0.13]*
	Two-year	0.05††† [0.04]	0.10††† [0.06]*	0.44††† [0.10]***	-0.05 [0.04]	-0.02† [0.05]	-0.08 [0.09]
	Three-year	-0.03† [0.04]	-0.05 [0.05]	0.31††† [0.09]***	-0.15 [0.04]***	-0.19 [0.06]***	-0.22 [0.09]**
	Four-year	0.02†† [0.05]	0.08†† [0.11]	0.48††† [0.14]***	-0.12 [0.06]**	-0.13 [0.09]	-0.14 [0.09]
	Five-year	-0.06 [0.04]	-0.09 [0.08]	0.35††† [0.11]***	-0.20 [0.04]***	-0.29 [0.06]***	-0.28 [0.10]***
	Six-year	-0.04 [0.05]	-0.10 [0.09]	0.31††† [0.12]***	-0.11 [0.04]***	-0.11 [0.06]*	-0.08 [0.09]
Time-invariant controls?		Yes	Yes	No	Yes	Yes	No
Time-varying controls?		Yes	Yes	Yes	Yes	Yes	Yes
Student Fixed Effects?		No	No	Yes	No	No	Yes
Nonprofit Schools		15	15	15	15	15	15
For-Profit Schools		5	5	5	5	5	5
Control Schools		71	71	71	71	71	71
Nonprofit Students		1,800–4,669	1,898–4,681	787–3,850	1,894–4,635	1,950–4,660	807–3,819
For-Profit Students		970–2,084	989–2,250	317–1,837	963–2,045	1,012–2,060	311–1,725
Control Students		9,843–24,194	9,912–24,535	5,196–21,860	9,851–23,927	9,993–24,002	5,229–21,595

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%; ††† (††) [†] indicates that the revoked for-profit effect is significantly different from the non-revoked nonprofit effect at 1% (5%) [10%]; robust standard errors adjusted for clustering within schools appear in brackets. Results for math and reading are estimated separately. Untreated schools include those in the bottom half of the (school aggregate) pre-treatment test-score distribution. Time-invariant controls include students' race and gender. Time-varying controls include students' special education and limited English proficiency status as well as indicator variables for whether the student made a structural or nonstructural move from the previous school year. In Model I, the dependent variable is the test-score level and prior-year test scores are included as controls. In Models II and III, the dependent variable is the decile-standarized test-score gain. All models include school fixed effects, grade-by-year fixed effects, and students' peer characteristics.

Table A5. Summary Statistics of Student Characteristics in 2002 and 2008

		All Students	Model I and II Students (at least one gain)	Model III Students (at least two gains)
2002	Number of Students	42,157	20,514	10,126
	% Special Ed	11.9%	12.4%	10.4%
	% Black	74.7%	74.3%	75.7%
	% Hispanic	14.3%	14.4%	14.7%
	Math Score	-0.22	-0.20	-0.26
	Reading Score	-0.20	-0.18	-0.24
2008	Number of Students	48,997	37,982	28,729
	% Special Ed	17.9%	20.4%	21.1%
	% Black	71.1%	71.4%	72.5%
	% Hispanic	18.8%	18.5%	18.1%
	Math Score	-0.21	-0.20	-0.20
	Reading Score	-0.21	-0.21	-0.21

Notes: Includes for-profit, nonprofit, and the lower-performing regular district schools. Math and reading scores are standardized by subject, test, grade, and year to have a mean of zero and standard deviation of one. The first column includes all students with a reading and math score available in the listed year. The second column includes all students with a reading and math test-score gain available in the listed year. The third column includes all students with a reading and math test-score gain available in the listed year and at least one other year.

Table A6. Summary Statistics of Student Characteristics in 2002, by Management

		Number of Students	Math Score	Reading Score
For-Profit	All Students	10,280	-0.40	-0.38
	Model I and II Students (at least one gain)	5,263	-0.39	-0.37
	Model III Students (at least two gains)	2,341	-0.49	-0.45
Nonprofit	All Students	3,919	-0.29	-0.27
	Model I and II Students (at least one gain)	1,905	-0.29	-0.25
	Model III Students (at least two gains)	801	-0.32	-0.30
Regular District	All Students	27,964	-0.14	-0.12
	Model I and II Students (at least one gain)	13,346	-0.12	-0.10
	Model III Students (at least two gains)	6,984	-0.17	-0.16

Notes: Math and reading scores are standardized by subject, test, grade, and year to have a mean of zero and standard deviation of one. "All Students" includes all students with a reading and math score available in the listed year. "Model I and II Students" includes all students with a reading and math test-score gain available in the listed year. "Model III Students" includes all students with a reading and math test-score gain available in the listed year and at least one other year.