Illusory Effects of Performance Management:

The Case of Contracts for Excellence in New York School Districts

Abstract

Externally imposed performance management systems often couple outcomes-based accountability with strong managerial reforms. While a rich literature documents the behavior of managers under performance management, literature on the direct performance effects of performance management systems is less conclusive. The Contracts for Excellence (C4E) program was a unique reform that contractually required 59 New York State public school districts to develop individualized plans based on "best practices" and report compliance and performance measures back to the state. This study evaluates C4E's impact on organizational performance using a difference-in-differences approach combined with propensity score matching. Findings point to negative or precisely estimated null effects on math and English test performance ranging from 0 to negative .14 standard deviations. Furthermore, the analysis uncovers suggestive evidence of undesirable institutional responses that may have compromised the reform, such as crowding out of local revenue collection and widespread inflation of performance metrics.

Introduction

During the last quarter of the twentieth century, public sector reformers renewed their search for improving efficiency in public sector organizations. Since public sector organizations lacked the economic incentives private sector organizations faced from market forces, they sought to implement other methods of incentivizing proper organizational management. This family of public sector reforms, known as New Public Management (NPM), was the source domain for the predominant accountability based reform of this era, performance management (Moynihan 2008). Performance management systems may be imposed both by external authorities, or they may be implemented internally by an organization. When performance management systems are imposed by external authorities, they often hold public managers accountable to rigorous performance standards, while also encouraging entrepreneurship by empowering managers with the discretion to lead broad organizational reforms (Moynihan 2006).

Performance management reforms have grown in popularity as a solution for the perceived problems of inefficiency and lack of accountability in public sector organizations. Early efforts such as the Government Performance Results Act of 1993 (GPRA), which stipulated that government agencies set measurable goals and submit annual performance reports, have sought to align the incentives of public managers with those of their external stakeholders (Kravchuk and Schack 1996). Empirical research on performance management often assumed that integration of performance information into managerial decision-making was the key determinant of managerial effectiveness, and a dominant stream of literature sought to understand the determinants of performance information use (Moynihan and Pandey 2010, Kroll 2015). Another stream of research explored "partial implementation" of performance

management systems, in which managers lack the managerial authority necessary to drive meaningful organizational change (Moynihan 2006, Nielsen 2013). The key assumption underlying both theoretical perspectives is that performance information should inform managerial decision-making – and that if managers successfully implement innovative reforms, organizational performance will improve.

An extensive empirical public management literature, reviewed by Kroll (2015), explores organizational decision-making under performance management, and a broader literature summarized by Heinrich and Marschke (2010) explores how dynamics of performance management reforms, such as employee motivation, system design and dynamic organizational responses shape reform effectiveness. Nonetheless, public management scholars have consistently expressed concerns about the limited base of empirical research that directly tests the relationship between performance information use and organizational performance, or the global effect of entire performance management systems on organizational performance. For instance, Boyne (2010, 217) concluded that "few researchers have sought to test the links between these activities [performance management] and organizational performance."

Kroll (2016) reiterated this argument, suggesting a framework to drive forward future research on the direct relationship between performance management and organizational performance. He reviews a growing body of public management research showing conflicting evidence with some positive results, though impacts are typically heterogenous by organizational characteristics and fidelity of reform implementation. Kroll argues that this heterogeneity is an inherent characteristic of performance management, and is driven by differences in organizational culture and implementation. Studies by Hvidman and Andersen (2013), Andersen (2008) and Nielsen (2013) likewise present conflicting views of the efficacy of performance

management in Danish public schools, with two finding null effects in public schools and the other finding effects of performance management that are mediated by managerial authority. A number of studies have now reported positive impacts of performance management (Walker, Damanpour, and Devece 2011, Carlson, Cowen, and Fleming 2013, Poister, Pasha, and Edwards 2013), but Kroll (2016) finds that the existence of performance management effects is dependent on an organization's managerial strategy. A recent meta-analysis by Gerrish (2016) synthesized a number of performance management related evaluation studies from outside of the public management literature, inferring small positive effects of performance management systems on organizational performance – but larger effects in studies using robust methodologies. Most of the studies in this meta-analysis either lacked experimental or quasi-experimental research design or did not offer a direct test of the relation between performance management and performance.

During the era of performance management, the field of public education was likewise transformed through its uptake of accountability-based reforms. In response to a perceived crisis in the competitiveness of American education, US public schools implemented a system of standardized testing and test-based accountability that developed over the course of the 1980's and 1990's. By 2001, every American state was required to implement a performance accountability system under the No Child Left Behind (NCLB) Act. Accountability reforms in public education offer an opportunity to evaluate the effects of performance management inspired reforms at a large scale. Both Carnoy and Loeb (2002) and Hanushek and Raymond (2005) evaluated pre-NCLB federal education reforms and found that states that implemented strong accountability systems experienced positive impacts on performance. Evaluation of NCLB was more nuanced, but the analysis of Dee and Jacob (2011) found some evidence of a

positive impacts on achievement. Other scholars found evidence of unintended consequences resulting from the incentives underlying NCLB's accountability system (Dee et al. 2010, Ladd and Lauen 2010, Reback, Rockoff, and Schwartz 2014). Ladd (2017) criticizes the NCLB reform for failing to empower schools with adequate resources or capacity building to lead the type of broad organizational reforms necessary to drive significant performance improvements.

A recent literature explores outcomes-based accountability and performance management systems in local education systems, particularly focusing on the case of New York City. In 2002, the New York City Department of Education instituted a series of organizational changes based on the philosophy that school administrators could improve performance if they were empowered to make decisions about their own schools. They offered principals the opportunity to sign accountability contracts in exchange for managerial independence. Under this system, schools were subjected to annual progress reports and principals were given greater autonomy over the budgeting, hiring, and all other decision-making processes previously controlled by the district (Childress et al. 2011, Destler 2016). Evaluation of this program, including quasi-experimental analysis (Wang and Yeung 2019), found qualified evidence of impacts on student achievement (Sun and Van Ryzin 2014). Several other studies have evaluated local and state-level accountability reforms, which on average show positive effects on school performance, but with unintended behavioral consequences (Figlio and Loeb 2011). Finally, a recent evaluation literature on "school turnaround" reforms, where external authorities mandate intensive managerial reforms in struggling schools to rapidly increase performance, has shown the potential for positive effects (Carlson and Lavertu 2018, Dee 2012), though these reforms are susceptible to failed implementation (Dragoset et al. 2017, Heissel and Ladd 2018).

The current study builds on this existing research through its assessment of a different form of performance management in New York State that tied increased resources to individualized managerial reform plans. As part of New York's Education Budget and Reform Act of 2007, which implemented a foundation aid funding formula aimed at closing socioeconomic performance gaps, the state introduced a program called Contracts for Excellence (C4E) which promised to deliver additional funding to 58 low performing districts in exchange for enhanced accountability and commitment to managerial reform.

C4E districts received noteworthy increases in state aid in the 2007-08 academic year, with the average district receiving an 11% increase, or a mean increase of $4.3 million for districts excluding New York City (NYC). In exchange for these funds, the districts were subject to outcome-based accountability measures and were required to annually create individualized plans for district-level management reform which documented how the new financial resources would be used. In the 2008-09 academic year, districts received another windfall increase, constituting a mean 26.4% increase in state aid over the two years, or an average of $9.7 million for non-NYC districts. In subsequent years, C4E districts did not receive additional aid increases, but their funding levels were maintained at the level of the 2008-09 academic year (New York State Education Department 2017).

In addition to providing revenues, the C4E program put into action a system for district-led managerial reform. In exchange for the financial investments, districts developed individualized management reform programs and were held accountable for improved and more equitable student outcomes. Figure 1 illustrates the major components of the C4E performance system, a framework we use to drive our theoretical questions and methodological considerations. Under C4E, districts selected from a menu of five evidence-based educational

reforms which they customized to construct individualized reform plans. These efforts were supplemented with up to 15% investment in experimental programs developed within the district. While all districts were working with the same managerial building blocks, the plans that resulted were unique to each district. For instance, while Buffalo City School District (CSD) highlighted its significant investment in class size reduction, Rochester CSD emphasized innovative programs such as district wide behavioral interventions to reduce suspensions and truancy. Syracuse CSD emphasized curriculum development, while Schenectady CSD used a significant portion of their funds to reopen an elementary school as a comprehensive early childhood development center. Thus, while district-level managerial discretion was partially constrained by best-practice guidelines, there was still wide variation in the programs that developed.

The state required school districts to submit revised plans at the beginning of each academic year, and to make the contracts public and to facilitate public comment on the proposal. At this point, plans were subject to approval from the state, and funding made contingent on successful execution of the plan. As the program progressed, select districts that demonstrated improved performance could graduate from the program, at which point they would no longer be subject to state and public scrutiny as a condition of their state aid dollars (See Figure 1). While the C4E cohort excluding NYC began with 55 school districts in 2007-08, the 2008-09 cohorts included only 38 districts. (In year two, three additional districts joined the program as well.) The following years had more districts graduating, with 31 remaining in 2009-10, 24 remaining in 2010-11 and 22 remaining in 2011-12. Districts that graduated from the program received the benefit of sustained state aid investments, but no longer had to comply with the oversight requirements of the C4E program (New York State Education Department 2017).

This study examines empirically the effectiveness of the C4E program in NYS. Using a seven-year panel of district-level data (2005-06 to 2011-12) on 650 NYS school districts, the study tests a number of hypotheses relating to the effects of the reform. The analysis first estimates treatment effects on organizational performance, and follows with examination of organizational measures that may speak to mechanisms of managerial decision-making. These subsequent analyses contribute suggestive evidence on how funding, management and accountability matter within performance management systems. Using a quasi-experimental difference-in-differences approach with district and year fixed effects, this study derives estimates of the effects of this program on organizational performance, and provides supplementary evidence relating to mechanisms. These findings contribute to a growing evidence base of studies testing the direct relationship between performance management and organizational performance, and can help scholars of public management develop a more nuanced understanding of when such reforms succeed, and when they fail.

Theory

The hypothesized effect of performance management systems on organizational performance is contested across different theoretical viewpoints. Some view performance management as misguided or detrimental to the morale and autonomy of public sector organizations. Radin (2006) argues that in performance management's narrow focus on linear measures of performance, it ignores the complexity of public organizations and the technical expertise of professionals. Frederickson and Frederickson (2006) argue that under New Public Management, resource-scare public organizations may lack the capacity to implement or respond to performance management systems, and that performance measures may fail to accurately communicate organizational outcomes. Further criticism of performance management systems

suggests that the measurement or organizational performance introduces potential for "gaming" or manipulation of reported outcomes (Hood 2006, 2012). Gerrish (2016) notes that, should performance management systems prove unsuccessful, policy-makers would shift their focus to less restrictive controls such as public service motivation and professional ethics (Perry and Wise 1990). While a number of literatures explore organizational responses to performance management (Kroll 2015) and how these responses and organizational dynamics shape the performance of performance management systems (Heinrich and Marschke 2010), a limited base of evidence tests directly the relationship between performance management and organizational performance (Boyne 2010, Kroll 2016).A richer understanding of the direct relationship between performance management and organizational performance provided by evaluation studies can help public management scholars balance criticism of these systems with their overwhelming appeal to policymakers.

Examining how the Contracts for Excellence performance management system operates over a longer time horizon, with salient measures of both institutional practices and organizational performance, presents an opportunity to contribute to this growing base of evidence. Of primary interest is whether the reform led to measurable increases in organizational performance. If C4E had a substantive institutional impact on school district performance, then one would expect to see translation into increased measures of organizational output. In the context of an educational organization, the most salient measure of performance is student math and reading proficiency. This leads to the primary hypothesis:

$\boldsymbol{H_1}$: *The C4E system directly improved organizational performance in the form of district-level student achievement outcomes.*

For the C4E program to translate into improved downstream performance, a prerequisite would be that accountability measures successfully introduce management reform and change institutional practices. Exploration of managerial mechanisms can help contextualize the overall performance findings. While the extent of managerial reform implementation or fidelity is unobservable in the available data, this study checks for program effects on a number of institutional measures that may provide suggestive evidence of managerial mechanisms. The following hypothesis addresses such institutional impacts:

$\boldsymbol{H_2}$: *The C4E performance management system caused school districts to change their behavior in a manner consistent with their management reform plans.*

The primary organizational incentive under C4E is to compel school districts to develop individualized management reform plans that tailor best-practice educational interventions and experimental treatments to the unique conditions of their district. Since every district was required to include class size reduction in their plans, measuring C4Es effect on class size can demonstrate whether C4E led to successful managerial reform. While this is by no means a perfect proxy for implementation or fidelity of managerial reforms, it does represent a managerial reform that every district was required to implement. To further explore mechanisms underlying the performance effects of the program, this study also examines the extent to which district resources and revenues change in response to the C4E program. Evaluation of the C4E program can provide a rigorous case study for how performance management systems affect performance. In contextualizing the impacts of the reform, the study seeks to provide further evidence relating to managerial mechanisms. Figure 1 suggests three main mechanisms that could theoretically differentiate the effectiveness of performance management reforms – money, management and accountability. To explain the performance effects of the C4E program, the

study includes a number of tests to explore how these mechanisms, which are not mutually exclusive, shape organizational practices and performance under C4E. These tests relate to the following hypotheses:

$\boldsymbol{H_{3A}}$: *The effects of the C4E program are attributable to the investment of financial resources in treatment districts.*

$\boldsymbol{H_{3B}}$: *The effects of the C4E program are attributable to more comprehensive accountability requirements.*

$\boldsymbol{H_{3C}}$: *The effects of the C4E program are attributable to individualized best-practice-based management reform.*

The C4E reform provided financial resources to targeted school districts and incentivized using those revenues towards reducing class size. As a first step, therefore, this study tests whether increased educational expenditures – which have been shown to improve student achievement (e.g. Jackson, Johnson and Persico, 2016; Lafortune, Rothstein and Schanzenbach, 2016) – drive program effectiveness ($\boldsymbol{H_{3A}}$**).** A second major component of the C4E program, common among performance management systems, is outcomes-based accountability measurement and tracking ($\boldsymbol{H_{3B}}$**)**. Prior literature has shown that accountability measures alone can improve performance in educational settings (Carlson, Cowen, and Fleming 2013). The third major component of the C4E program was individualized managerial reforms, which included both mandated changes such as class size reductions, but also allowed for increased programmatic experimentation and managerial discretion across districts ($\boldsymbol{H_{3C}}$**)**. To assess the contribution of each of these factors, the analysis employs a series of modifications to the main model (described fully in the results section) to provide suggestive evidence of the different factors at play.

The C4E program not only intended to improve outcomes for the general student population; it also emphasized a significant equity component. It required that districts target resources towards underperforming groups, such as economically disadvantaged students and students with disabilities. Critics of performance management systems like C4E often worry that they can have uneven distributional impacts, as organizations favor clients with the greatest likelihood of improvement, and neglect disadvantaged individuals (Heckman and Smith, 1997). These considerations lead to the fourth hypothesis, which is tested using performance data specific to economically disadvantaged student populations:

$\boldsymbol{H_4}$: *The C4E system enhanced measures of vertical equity by raising the performance of disadvantaged student groups more so than corresponding increases for the general student population.*

Data

The analysis in this paper is based on publicly available data from the New York State Education Department (NYSED). Financial variables are obtained from the Fiscal Analysis and Research Unit's (FARU) Fiscal Profile Reporting System (FPRS). Test score outcomes, student demographics and other control variables are drawn from the NYSED School Report Cards. Merging these data sources together results in a complete match of 677 major school districts. Because some of these districts only serve elementary or high school students, the analysis sample includes only the 650 K-12 school districts in NYS. The sample excludes the New York City Public School District, because it differs dramatically from the rest of the state on size, population, and organizational and financial structure. This dataset spans seven years, from the 2005-06 to 2011-12 academic years.

The current study analyzes the effect of C4E on two sets of dependent variables. The first set of dependent variables captures organizational resources and resource allocation. Student-teacher ratio (student enrollment divided by the total number of teachers) and per-pupil expenditures (PPE) (total expenditures divided by enrollment) measures attempt to operationalize managerial change resulting from the program. Local revenue per-pupil (total local revenue divided by enrollment) and state-aid per-pupil (total state-aid revenue divided by enrollment) reflect the source of these institutional resources.

The second set of dependent variables includes performance data from standardized math and English/Language Arts (ELA) tests delivered annually to students in grades 3 through 8. New York's standardized testing program begins in grade 3 and ends in grade 8, with high school students assessed using a different and less uniform system. The final measure consists of average performance of students across grades 3-8. The study also employs equivalent measures of academic performance (grades 3-8 math and English) that pertain to only economically disadvantaged students. This is possible because New York State school districts not only report average test scores for their entire student body, but also average test scores pertaining to a number of student subgroups. All performance measures are standardized to have mean zero and standard deviation of one by school year, but raw scaled test scores are also considered.

This study uses a binary policy indicator that equals one if the district participated in C4E and the time period is post 2007-08 school year implementation ($Treatment * Post$), and zero otherwise. Though some districts departed the original C4E cohort as the program progressed, at which point they were no longer subject to accountability requirements, districts are not removed from the policy treatment group as they leave in the baseline model. This specification assumes that C4E funding increases and the effects of management reforms, if not the reforms

themselves, persisted after graduation, and that the accountability measures of the program contributed minimally to explaining variation in program outcomes. These assumptions are relaxed in an alternative model, described in the analytical section of this text. In addition, the sample includes three districts that entered the program one year late in 2008-09. The $Treatment * Post$ indicator is modified to be coded zero during the 2007-2008 academic year for these 3 districts.

Finally, the dataset includes a rich set of covariates incorporating demographic measures and district-level institutional characteristics. These include percent of students eligible for free lunch (a proxy for low-income status), percent of students from a racial/ethnic minority group, percent of students with limited English proficiency (LEP), percent of students with disabilities, student enrollment, debt payments per pupil and average teacher salary. All financial variables are adjusted for inflation, and reported in year 2016 dollars. Other minor details from the variable construction process are provided in an appendix.

Summary statistics of all dependent and independent variables are shown in Table 1. Of note, districts on average receive just over $8,000 per pupil in state aid and approximately $10,000 in local revenue (NYS leads the nation in highest educational expenditures). Approximately 6% of the sample are C4E districts. Table 2 presents summary statistics comparing characteristics of C4E and non-C4E districts in pre-treatment period, with a column of p-values for means comparison tests for each variable. These comparisons illustrate that C4E districts tend to be significantly larger, more diverse, more disadvantaged, and lower-performing than non-C4E districts.

Methods

One could simply estimate the association between Contracts for Excellence and student and institutional performance using ordinary least squares (OLS) regression. However, because participation in the C4E program was not randomly assigned, but rather mandated to low-performing districts, C4E participation is likely to be correlated with unobserved characteristics of participating districts. It is possible to address the endogenous nature of the C4E reform by specifying a generalized difference-in-differences model according to the following model:

$$y_{dt} = \gamma_0 Post \times Treatment_{dt} + \gamma_1 X_{dy} + \theta_d + \tau_t + \varepsilon_{dt} \text{ (Equation 1)}$$

In this equation, Y_{dt} is an outcome of interest for district d in year t, X_{dt} is a vector of district level demographic and institutional characteristics, and ε_{dt} is a stochastic error term for district d in year t. The measure $Post_t$ equals one if year t is after C4E implementation and zero otherwise, $Treatment_d$ equals one if district d is at any point in time identified as a C4E district and zero otherwise, and $Post \times Treatment_{dt}$ is the interaction of the two variables. Including the district fixed effects θ_d accounts for all time-invariant characteristics of C4E districts that distinguish them from other districts, and including year fixed effects τ_t accounts for all unobserved characteristics of the post-implementation period. This functional form mitigates the selection bias problem and allows β_3 – the coefficient of interest – to capture arguably causal effects of the C4E reform.

All models are estimated with Huber-White robust standard errors clustered by district, to address heteroscedasticity and autocorrelation within districts. As is the case with all difference-in-differences approaches, the empirical strategy relies on the parallel trends assumption: that C4E district pre-treatment trends were similar to those of non-C4E districts. The results section includes a series of placebo tests to investigate this assumption.

Another potential concern is that C4E districts differ systematically from other districts in ways that could drive differential trends in student performance in the post-reform period. For this reason, the analysis includes estimates of the model described above which restrict the control group to a smaller set of districts matched one-to-one with C4E districts based on a series of baseline academic, financial, and demographic district characteristics. To select this sample, propensity scores were estimated using a probit regression restricted to the last pre-treatment year (2007) and including all available covariates listed in Table 1. The propensity score regression is included in Appendix Table 1. Each C4E district was then matched to its nearest neighbor without replacement, and absorbed the entire panel of each match to form a comparison group. This is the preferred model as it allows estimation of treatment effects compared to a similar comparison group, against whom one should expect to see a genuine effect. Table 3 provides a comparison of descriptive statistics between C4E and matched districts. While the predominantly poor and urban composition of the C4E treatment group drives residual imbalance on a number of measures including academic performance, poverty and minority composition, the differences are greatly attenuated compared to the full sample. In the results in the following section, the models estimated against the full sample are referred to as model 1 and those estimated against the matched sample as model 2.

Results

Effects on Academic Outcomes

To assess the impact of the C4E reform on performance, estimates are provided from the two main models for standardized measures of average scores in grade 3-8 math and English end-of-year exams. These results, shown in Table 4, reveal a precisely-estimated zero effect on math in model 1 and negative effect of .12 standard deviations (SDs) in model 2 that is

significant at the .10 level. The English results reveal a more troubling impact. Negative program impacts of approximately .06 SDs in model 1 (statistically significant at the .10 level) and of a larger .16 SDs in model 2 (significant at the .05 level) are observed. This suggests that the C4E program had either a null or negative impact on organizational performance.

This finding presents a puzzle. The finding that C4E districts created null or negative achievement gains, especially when compared with similar districts, clashes with the purported success of the program. The majority of C4E districts graduated from the program, most within the first two years, indicating that they were at least able to convince policy-makers that their performance was improving.

To reconcile these positive perceptions of the program with the more pessimistic empirical findings, trends in the unstandardized mean test scores during the period of this study are considered. The graphs of these trends are included in figures 2 and 3. These figures break down the trends into four separate groups, consisting of the top half of the achievement distribution, C4E program graduates, C4E districts that did not graduate (persisters), and control districts in our propensity score matched sample. The trends in math show strong growth in average test scores among all districts. The English scores show a more gradual upward trend, except among high-performing districts, indicating possible C4E gains relative to the top of the achievement distribution. Gains are especially conspicuous among all groups in the first two years of the program (2008 and 2009), when nearly half of C4E districts, and three-quarters of C4E graduators, graduated from the program. This goes half way to explaining the discrepancy between our results and the apparent success of the program in graduating districts. Because the models use annually-standardized performance measures, they capture gains over time relative to

other districts – of which there were none – rather than absolute gains which would have looked more promising.

The study further considers whether these large gains in unstandardized test scores were the result of test score inflation as districts adapted to New York State's standardized testing regime. The sharp increases in test scores during this period were documented in the New York Times, where they attracted significant criticism from experts who warned that the improvements were too good to be true. This suspicion was confirmed in 2010, when NYS responded to criticism by increasing the difficulty of the exams, and student performance plummeted (Medina 2010) (a result that can be observed in Figures 3 and 4). In studies of nationwide accountability systems, Hanushek and Raymond (2005) and Dee and Jacob (2011) estimate treatment effects on National Assessment of Educational Progress (NAEP) exam performance, because this is a low-stakes exam that is not connected to accountability measures and therefore not susceptible to gaming responses or rapid inflation. When consulting trends in NAEP scores over the period of the study, there is no indication of an upward trend for New York schools in any performance quartile (NAEP 2018). This suggests that the performance measure gains experienced by NYS schools during this period were not a result of true performance gains, but rather from unintended responses of schools or districts to the measures themselves.

These findings support the conclusion that test score inflation during the period of the study allowed C4E districts to post positive performance growth each year, while actually losing ground relative to similar districts. These illusory gains allowed them to evade accountability requirements, which confounded the ability of policy-makers to effectively manage the C4E program. To illustrate how this confounding could occur, it is possible to estimate treatment

effect models with the un-standardized test score measures displayed in Figures 3 and 4. The results of these analyses are included in Table 5. When judged against the entire sample of NYS school districts (model 1), these models estimate large statistically significant positive effects, a result which would have certainly pleased the state education department. However, when observing the effect relative to the matched sample of similar districts (model 2), the positive estimates vanish.

Effects on Organizational Investments

To help contextualize the negative performance effects of the C4E program, the analysis next estimates the effects of C4E on organizational responses to determine if C4E had the intended effect on operations within school districts. The two organizational measures of interest are student-teacher ratio and PPE. Because the C4E program required managerial reform plans that included a class size reduction component, treatment effects on the student-teacher ratio measure may serve as a proxy indication of implementation of the targeted managerial reform. Since the program promised increased revenues to treated districts, the PPE measure likewise provides an indication of whether these revenues translated into increased organizational investments. While these measures are not perfect proxies for managerial implementation or fidelity, they may provide suggestive evidence in the absence of observable measures of managerial decision-making. Since C4E contracts all required reduced class sizes and increased investment in innovative programs, districts which successfully implemented their reforms should demonstrate changes in these measures.

Table 6 provides the results of these analysis. In both model 1 and model 2, C4E districts did not decrease their student-teacher ratios. This result indicates the incentivized managerial reforms under C4E simply may not have occurred, which could explain the failed performance

effects of the program. The results of these analyses examining treatment effects on educational expenditures, also in Table 4, are again concerning. C4E treatment districts actually reported lower PPE by approximately $600 in model 1, a relationship that is statistically significant at the .05 level. The model 2 estimate is a decrease of approximately $160 as compared to the matched group, though the effect is statistically indistinguishable from zero. Nonetheless, a null negative finding for a resource metric that should have increased under the treatment is still surprising.

To better understand these findings, treatment effects are estimated on two revenue categories, local revenue per pupil and state aid revenue per pupil, which each make up approximately half of district revenue. The results of these analyses (Table 7) show that C4E districts collected locally approximately $900 per pupil less than all other districts following implementation of the program (model 1), and approximately $600 less in model 2. It seems that the promise of increased state-aid under C4E led districts to engage in less independent revenue collection. The results for state-aid show that treatment districts received approximately $400 in increased state aid revenue (significant at .01 level in both models). Therefore, treatment districts reduced local revenue collection by approximately $2 for every $1 they received in additional aid. This is consistent with a literature on "crowding out" that documents school districts responding to intergovernmental grants by decreasing local revenue (Cascio, Gordon, and Reber 2013, Gordon 2004). This crowding-out phenomenon provides a credible explanation for the lack of incentivized organizational responses (smaller class sizes) under the C4E program.

Analyses of Mechanisms

The analyses in the preceding sections suggest that the C4E program led to perverse organizational responses, incentivizing treated districts to collect less revenue, leading to decreased resources overall and possibly a lack of managerial change. Furthermore, the

performance analysis suggests that gaming behaviors led to rapid test score inflation, which falsely created the perception of positive performance growth and confounded the accountability mechanisms of the program. After correcting for this inflation, the program shows either precisely estimated null impacts, or possible negative effects. The original hypotheses H_{3A}, H_{3B}, and H_{3C} put forth three possible mechanisms through which the C4E performance management reform could have improved student performance: financial resources, accountability requirements, and best-practice-based management reform. Based on the results that C4E did not decrease class size, and that the increased state funding merely substituted for corresponding decreases in local revenues, one can conclude that neither financial resources (H_{3A}) nor best-practice-based managerial reforms (H_{3C}) positively contributed to student performance. However, it is possible that decreased resources or poorly executed managerial reforms led to negative impacts. Furthermore, whether the accountability requirements of the C4E reform had an impact on district performance (H_{3B}) requires further investigation.

Appendix Table 7 includes estimates from models that control for per-pupil expenditures. Because C4E led to lower resources in treated districts, it is important to understand whether there were truly null or negative impacts of the performance management reform itself, or whether possible positive effects are masked by the countervailing reduction in resources. The findings after controlling for spending show that any program impacts do not result from changes in financial resources available to districts. Estimates are essentially unchanged in terms of effect size or significance. While these analyses should be interpreted as descriptive in nature given the endogeneity of school spending, they provide suggestive evidence against the interpretation that lack of performance improvements resulted from the failure of targeted aid to carry over to increased expenditures.

To check specifically for potential impacts of the instituted accountability system, the analysis includes estimates from models that leverage variation among treated districts in length of exposure to oversight under C4E. Because some districts graduated from the program, it is possible to compare performance in treated districts both during and after the period of oversight, with the assumption that increased funding and managerial reforms would have persisted but accountability would have ended. To do so the analysis introduces an indicator of exposure to oversight that is coded one if a treatment district was currently subject to accountability oversight and zero otherwise. When added to the main model, this variable creates a "triple-differences" approach to estimate heterogenous impacts of the C4E program by oversight status. The results of these analyses, included in Appendix Table 8, confirm no evidence of heterogenous impact by accountability status, with precisely estimated null coefficients on the accountability indicator for both test score measures. The failure of these models to detect performance impacts of the accountability reforms under C4E is consistent with the hypothesis that secular trends in the test score measures confounded accountability mechanisms.

This lack of organizational change in response to accountability pressure helps to explain the overall null effects of the program, along with the inflationary trends in raw test scores observed in the academic results section. However, the precise null effects revealed in these analyses do not suggest that thwarted accountability measures in any way worsened impacts on those districts differentially exposed to longer accountability periods. If one believes the negative treatment effects observed in some of the main analyses, the lack of the performance effects assumed by H_{3A} and $H_{3B,}$ leaves H_{3C} as the only remaining explanation for adverse effects by a process of elimination. Under this assumption, a possible explanation for negative consequences of C4E would be that managerial reforms under the program were misguided, disruptive, poorly

executed or diverted resources from core mission objectives. Failures due to poor implementation or administrative burdens have become a common finding in evaluations of performance-based educational reforms such as school turnarounds (Heissel and Ladd 2018, Dragoset et al. 2017), and similar mechanisms could have been in play under the C4E reform. However, without observable measures of managerial implementation, it is difficult to make strong conclusions about the managerial aspects of the program.

The final hypothesis concerns equity effects of the C4E program. The analysis corresponding to this hypothesis checks whether the possibly negative impacts observed in the main models are driven by worse performance among disadvantaged student groups, or whether the null effects mask better or worse outcomes among vulnerable student populations. To do this, the study provides estimates from models with average test score data that only includes economically disadvantaged students. The results of these analyses are included in appendix table 9. The results are less precisely estimated, but significantly larger in magnitude, than from the main analyses, with very large negative effects observed in English scores in Model 2, which are significant at the .05 level. The large size of these coefficients suggests that negative effects on economically disadvantaged students were contributing to the negative effects observed in the full sample of students, and perhaps driving them in English. This raises significant equity concerns about the possibility that performance management reforms have disproportionate adverse impacts on disadvantaged clients when improperly implemented.

Robustness Checks

The difference-in-differences models used in this analysis assume that outcomes in the treatment and comparison group would have been the same in the absence of treatment, an assumption commonly referred to as parallel trends. A “placebo test,” in which one re-specifies

the empirical model using only the two pre-treatment periods and assigns an “artificial” treatment to C4E districts in the year before implementation occurs, can test this assumption. According to Mora and Reggio (2017), this test may be used to verify parallel trends assumptions in difference-in-difference models with two pre-treatment periods. This allows one to determine whether C4E districts displayed differential trends in the dependent variables during the pre-treatment period (which could indicate violation of the parallel trends assumption). The results of these analyses are included in appendix tables 3 through 6 and 10, each one corresponding to a main table presented in the paper.

Overall, none of the placebo treatments have statistically significant effects on any institutional outcomes or organizational performance within our preferred model specifications. This allows one to conclude more confidently that the C4E program lacked any true positive impacts on either targeted managerial reforms or on student performance. The only clear violation of pre-treatment trends occurs when testing the unstandardized version of math and reading scores (see appendix table 6), which is consistent with the divergent achievement trends issue we identified and discussed earlier from figures 2 and 3.

Discussion

The results of this analysis construct a new narrative of the ways in which institutional gaming responses to performance management regimes can thwart true performance improvement. The study presents evidence that when NYS school districts were offered extra resources in exchange for accountability and commitment to managerial reform, they responded by substituting away from local revenue collection, severely reducing their institutional resources. This led to a failed managerial reform that did not produce measurable changes in targeted institutional outcomes. The analysis of the performance trends of C4E districts

compared to other similar school districts provides indications of how this was allowed to happen even in the context of a rigorous accountability system. Low-performing districts rapidly adapted to the performance measures they were held accountable for, leading to false positive performance gains that fooled the accountability mechanisms in place. After correcting the analysis for this widespread test inflation, the C4E program has null or negative effects on math and English performance, ranging from 0 to negative .14 standard deviations.

To help contextualize the null or negative performance effects of the reform, supplemental analyses explored possible causal mechanisms operating within the C4E performance management system, including money, accountability and management. After controlling for the revenue substitution triggered by this program, the results suggest that resource deficits did not mask positive performance effects of the C4E program. Models exploiting variation in exposure to oversight under the program further suggest that the accountability mechanisms under C4E were ineffectual and did not contribute to changes in performance. By a process of elimination (according to the causal logic specified in Figure 1), this may indicate that residual negative impacts on math and English performance measures were related to managerial changes under the program, possibly due to organizational disruption resulting from the poorly-resourced managerial reforms.

While the original objective of the study was to simply test the efficacy of this performance management reform, a more interesting series of findings emerged about institutional responses to a performance management reform. The findings should not be interpreted as concluding that performance management systems in general cannot boost performance. Instead, the lack of improvement uncovered in this study suggests a number of ways in which performance management systems can go wrong. First, providing resource

incentives within performance management systems can fail, as it creates a moral hazard for public organizations to decrease effort in independent revenue collection. Second, the findings provide empirical evidence of a common criticism of performance measures, that they incentivize public organizations to "hit the target and miss the point" (Hood 2006, 2012) by gaming metrics without improving the quality of public service delivery. Finally, the supplemental analyses suggest that accountability failures and ineffective managerial change, rather than resource deficits, were the drivers of this failed performance management system.

Beyond theoretical contributions, this study makes a more technical contribution to the study of performance effects in the public management literature. We document that a common performance measure, end of year exam scores, is subject to compromise by secular trends. This can lead to distributional changes that produce false positives and defy even standard quasi-experimental econometric methods. For this reason, these findings emphasize that all performance metrics should be compared to measures that are standardized annually relative to the performance of other districts, and, when possible, comparisons should be made to performance measures that are not susceptible to gaming for accountability purposes, such as the NAEP test. This study also demonstrates that using a difference-in-difference approach applied to a propensity-score-matched sample, instead of to the full sample, can protect researchers from estimating false positives under circumstances of likely parallel trend violations.

Taken in sum, the experience of NYS school districts under C4E highlights the long-recognized importance of incentives built in to performance measurement systems. If stakeholders are not careful in their design of performance systems, they can lead to outcomes, such as decreased revenue collection or performance measure gaming, that prevent public organizations from enacting true managerial improvements. This was the case under C4E, where

a highly-publicized and expensive reform failed to generate performance gains. While this case does not rule out the possibility that performance management systems can work, it does underscore the various pitfalls that can thwart successful implementation.

References

Andersen, Simon Calmar. 2008. "The Impact of Public Management Reforms on Student Performance in Danish Schools." *Public Administration* 886 (2541-558).

Boyne, George A. 2010. "Performance management: Does it work? ." In *Public management and performance: Research directions*, edited by Richard M. Walker, George A. Boyne and Gene A. Brewer, 207-226. New York: Cambridge University Press.

Carlson, Deven E., Joshua M. Cowen, and David J. Fleming. 2013. "Third-Party Governance and Performance Measurement: A Case Study of Publicly Funded Private School Vouchers." *Journal of Public Administration Research and Theory* 24 (4):897-922.

Carlson, Deven E., and Stephane Lavertu. 2018. "School Improvement Grants in Ohio: Effects on Student Achievement and School Administration." *Educational Evaluation and Policy Analysis* 40 (3):287-315.

Carnoy, Martin, and Susanna Loeb. 2002. "Does external accountability affect student outcomes? A cross-state analysis." *Educational Evaluation and Policy Analysis* 24 (4):305-331.

Cascio, Elizabeth U., Nora Gordon, and Sarah Reber. 2013. "Local Responses to Federal Grants: Evidence from the Introduction of Title 1 in the South." *American Economic Journal: Economic Policy* 5 (3):126-159.

Childress, Stacey, Monica Higgins, Ann Ishimaru, and Sola Takahashi. 2011. "Managing for results at the New York City Department of Education." In *Education reform in New York City: Ambitious change in the nation's most complex school system*, edited by J. O'Day, C. Bitter and L.M. Gomez. Cambridge, MA: Harvard Education Press.

Dee, Thomas S. 2012. "Title." NBER Working Papers.

Dee, Thomas S., and Brian A. Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30 (3):418-446.

Dee, Thomas S., Brian A. Jacob, Caroline M. Hoxby, and Helen F. Ladd. 2010. The impact of No Child Left Behind on students, teachers and schools. In *Brookings papers on economic activity*. Washington D.C.: Brookings Institution.

Destler, Katharine Neem. 2016. "Creating a performance culture: Incentives, climate and organizational change." *American Review of Public Administration* 46 (2):201-225.

Dragoset, Lisa, Jaime Thomas, Mariesa Herrmann, John Deke, Susanne James-Burdumy, Cheryl Graczewski, Andrea Boyle, Rachel Upton, Courtney Tanenbaum, and Jessica Giffin. 2017. School Improvement Grants: Implementation and Effectiveness. U.S. Department of Education.

Figlio, David, and Susanna Loeb. 2011. "School Accountability." In *Handbook of the Economics of Education*, edited by Eric A. Hanushek, Stephen Machin and Ludger Woessman, 383-421. The Netherlands: North-Holland: Elsevier.

Frederickson, David G., and H. George Frederickson. 2006. *Measuring the Performance of the Hollow State*. Washington, DC: Georgetown University Press.

Gerrish, Ed. 2016. "The impact of performance management on performance in public organizations: A meta-analysis." *Public Administration Review* 76 (1):48-66.

Gordon, Nora. 2004. "Do Federal Grants Boost School Spending? Evidence from Title1." *Journal of Public Economics* 88:1771-1792.

Hanushek, Eric A., and Margaret E. Raymond. 2005. "Does school accountability lead to improved student performance?" *Journal of Policy Analysis and Management* 24 (2):297-327.

Heinrich, Carolyn, and Gerald Marschke. 2010. "Incentives and their Dynamics in Public Sector Performance Management Systems." *Journal of Policy Analysis and Management* 29 (1):183-208.

Heissel, Jennifer A., and Helen F. Ladd. 2018. "School Turnaround in North Carolina: A Regression Discontinuity Analysis." *Economics of Education Review* 62 (302-320).

Hood, Christopher. 2006. "Gaming in Targetworld: The Targets Approach to Managing British Public Services." *Public Administration Review* 66 (4):515-521.

Hood, Christopher. 2012. "Public Management by Numbers as a Performance-Enhancing Drug: Two Hypotheses." *Public Administration Review* 72 (S1):585-592.

Hvidman, Ulrik, and Simon Calmar Andersen. 2013. "Impact of Performance Management in Public and Private Organizations." *Journal of Public Administration Research and Theory* 24 (1).

Kravchuk, Robert S., and Ronald W. Schack. 1996. "Designing effective performance measurement systems under the Government Performance Results Act of 1993." *Public Administration Review* 56 (4):348-358.

Kroll, Alexander. 2015. "Drivers of Performance Information Use: Systematic Literature Review and Directions for Future Research." *Public Performance & Management Review* 38 (3):459-486.

Kroll, Alexander. 2016. "Exploring the Link Between Performance Information Use and Organizational Performance: A Contingency Approach." *Public Performance & Management Review* 39:7-32.

Ladd, Helen F. 2017. "No Child Left Behind: A Deeply Flawed Federal Policy." *Journal of Policy Analysis and Management* 36 (2):461-469.

Ladd, Helen F., and Douglas L. Lauen. 2010. "Status versus growth: The distributional effects of school accountability policies." *Journal of Policy Analysis and Management* 29 (3):426-450.

Medina, Jennifer. 2010. "On New York School Tests, Warning Signs Ignored." *The New York Times*, October 10, 2010.

Mora, Ricardo, and Iliana Reggio. 2017. "Alternative diff-in-diffs estimators with several pre-treatment periods." *Econometric Reviews* (forthcoming).

Moynihan, Donald P. 2006. "Managing for Results in State Government: Evaluating a Decade of Reform." *Public Administration Review* 66 (1):77-89.

Moynihan, Donald P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*. Washington, DC: Georgetown University Press.

Moynihan, Donald P., and Sanjay K. Pandey. 2010. "The Big Question for Performance Management: Why Do Managers use Performance Information?" *Journal of Public Administration Research and Theory* 20 (4):849-866.

NAEP. 2018. "State Profiles." https://www.nationsreportcard.gov/profiles/stateprofile?chort=1&sub=MAT&sj=&sfj=NP&st=MN&year=2017R3.

New York State Education Department. 2017. Contracts for Excellence.

Nielsen, Poul A. 2013. "Performance Management, Managerial Authority, and Public Service Performance." *Journal of Public Administration Research and Theory* 24 (2):431-458.

Perry, James L., and Lois Recascino Wise. 1990. "The motivational bases of public service." *Public Administration Review* 50 (3):367-73.

Poister, Theodore H., Obed Q. Pasha, and Lauren Hamilton Edwards. 2013. "Does Performance Management Lead to Better Outcomes? Evidence from the U.S. Public Transit Industry." *Public Administration Review* 73 (4):625-636.

Radin, Beryl A. 2006. *Challenging the Performance Movement: Accountability, Complexity and Democratic Values*. Washington, D.C.: Georgetown University Press.

Reback, Randall, Jonah Rockoff, and Heather L. Schwartz. 2014. "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under No Child Left Behind." *American Economic Journal: Economic Policy* 6 (3):207-241.

Sun, Rusi, and Gregg G. Van Ryzin. 2014. "Are performance management practices associated with better outcomes? Empirical evidence from New York public schools." *American Review of Public Administration* 44 (3):324-338.

Walker, Richard M., Fariborz Damanpour, and Carlos A. Devece. 2011. "Management innovation and organizational performance: The mediating effect of performance management." *Journal of Public Administration Research and Theory* 21 (2):367-386.

Wang, Weijie, and Ryan Yeung. 2019. "Testing the Effectiveness of "Managing for Results": Evidence from an Education Policy Innovation in New York City." *Journal of Public Administration Research and Theory* 29 (1):84-100.

Tables and Figures

Figure 1. Diagram of Contracts for Excellence (C4E) Performance Framework and Theoretical Hypotheses

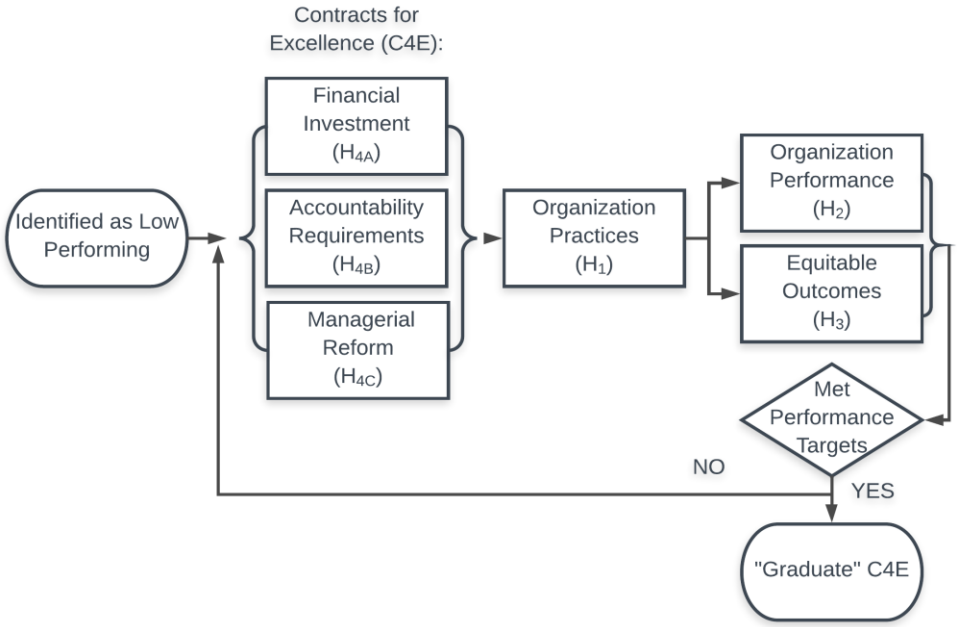

Figure 2. Histogram of District Average 8th Grade Math Scores: 2006 and 2013 School Years

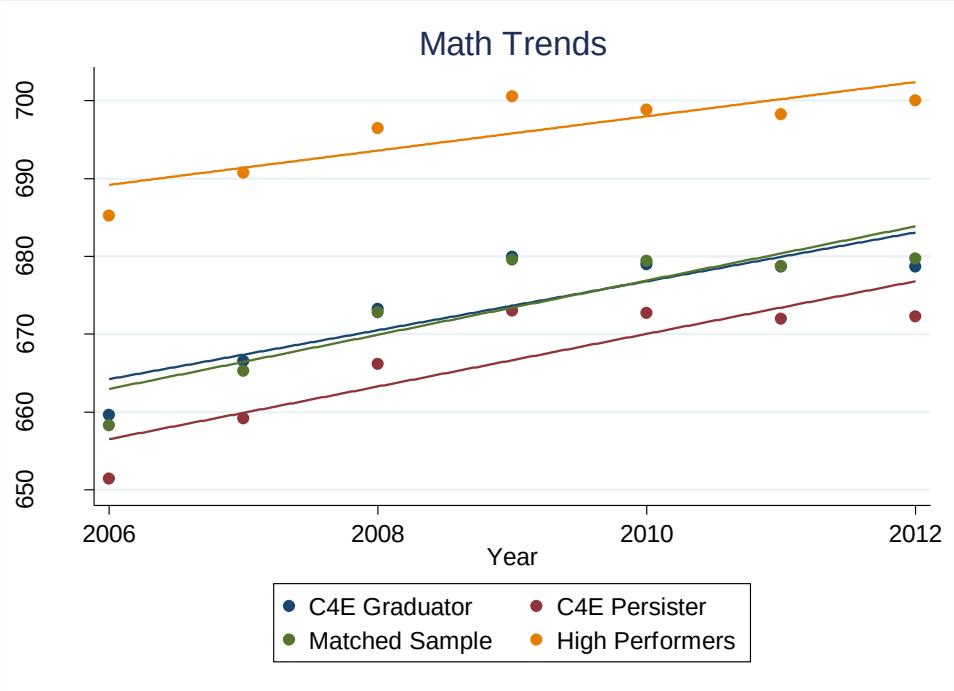

Figure 3. Histogram of District Average 8th Grade ELA Scores: 2006 and 2013 School Years

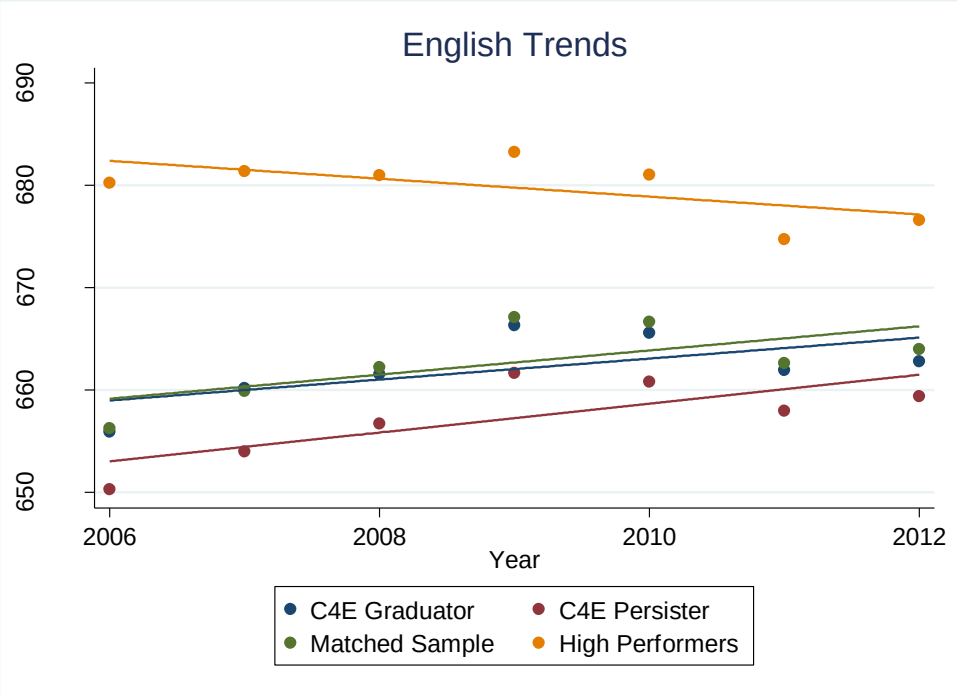

Table 1: Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Math	4489	679.88	12.22	626.50	717.00
English	4489	667.72	8.71	630.33	696.33
Math (Standardized)	4489	0.00	1.00	-3.58	3.41
English (Standardized)	4489	0.00	1.00	-3.46	3.37
Math (Economically Disadvantaged, Standardized)	3929	0.00	1.00	-4.21	4.41
English (Economically Disadvantaged, Standardized)	3929	0.00	1.00	-3.94	4.43
C4E Treatment	4489	0.09	0.29	0.00	1.00
Treatment*Post	4489	0.06	0.24	0.00	1.00
Teacher-Student Ratio	4489	11.57	1.88	4.21	30.77
PPE	4489	20.95	4.87	12.16	83.06
Local Revenue PP	4489	10.66	6.95	0.90	73.80
State Aid PP	4489	8.33	3.90	0.93	23.46
Debt PPE	4489	1.63	1.23	0.00	47.09
Teacher Salary	4489	77.03	20.17	38.08	147.28
Enrollment	4489	2654.17	3347.77	99.00	43436.00
% Free Lunch	4489	22.59	14.76	0.00	94.00
% Minority	4489	15.48	19.76	0.00	100.00
% LEP	4489	1.92	3.99	0.00	33.00
% SWD	4489	15.30	4.22	0.51	56.28

Table 2: Comparison of C4E and Non-C4E Districts			
Factor	Non-C4E	C4E	p-value
N	1145	116	
Math	670.81 (11.75)	660.31 (11.26)	<0.001
English	666.05 (10.53)	655.94 (9.52)	<0.001
Math (Standardized)	0.09 (0.97)	-0.83 (0.92)	<0.001
English (Standardized)	0.09 (0.97)	-0.87 (0.87)	<0.001
Math (Economically Disadvantaged, Standardized)	0.08 (0.96)	-0.71 (1.05)	<0.001
English (Economically Disadvantaged, Standardized)	0.09 (0.96)	-0.77 (0.99)	<0.001
Teacher-Student Ratio	12.12 (2.34)	13.04 (1.60)	<0.001
PPE	19.65 (4.64)	18.14 (4.88)	<0.001
Local Revenue PP	10.17 (6.77)	7.82 (5.75)	<0.001
State Aid PP	7.63 (3.56)	8.15 (2.80)	0.13
Debt PPE	1.42 (0.76)	1.11 (0.64)	<0.001
Teacher Salary	77.88 (19.55)	78.81 (17.94)	0.62
Enrollment	2322.63 (2170.26)	6708.86 (7836.46)	<0.001
% Free Lunch	18.65 (12.48)	34.11 (16.72)	<0.001
% Minority	12.40 (17.17)	32.65 (28.32)	<0.001
% LEP	1.57 (3.19)	5.28 (7.67)	<0.001
% SWD	15.26 (4.08)	17.88 (6.22)	<0.001

Table 3: Comparison of C4E and PSM Matched Comparison Districts			
Factor	PSM Matched	C4E	p-value
N	113	116	
Math	663.51 (11.93)	660.31 (11.26)	0.038
English	659.32 (10.09)	655.94 (9.52)	0.010
Math (Standardized)	-0.56 (0.99)	-0.83 (0.92)	0.035
English (Standardized)	-0.56 (0.94)	-0.87 (0.87)	0.010
Math (Economically Disadvantaged, Standardized)	-0.44 (1.09)	-0.71 (1.05)	0.057
English (Economically Disadvantaged, Standardized)	-0.44 (0.97)	-0.77 (0.99)	0.010
Teacher-Student Ratio	12.84 (2.61)	13.04 (1.60)	0.48
PPE	18.62 (3.32)	18.14 (4.88)	0.39
Local Revenue PP	8.60 (4.16)	7.82 (5.75)	0.24
State Aid PP	7.95 (3.00)	8.15 (2.80)	0.61
Debt PPE	1.23 (0.86)	1.11 (0.64)	0.22
Teacher Salary	78.11 (18.71)	78.81 (17.94)	0.77
Enrollment	4325.22 (3456.51)	6708.86 (7836.46)	0.003
% Free Lunch	27.33 (15.17)	34.11 (16.72)	0.001
% Minority	24.27 (27.02)	32.65 (28.32)	0.023
% LEP	3.93 (6.07)	5.28 (7.67)	0.14
% SWD	17.08 (3.66)	17.88 (6.22)	0.24

Table 4: Effects of C4E on Academic Outcomes (Standardized)				
	(1)	(2)	(3)	(4)
VARIABLES	Math	Math	English	English
Treatment*Post	-0.02	-0.11+	-0.06+	-0.14*
	(0.05)	(0.07)	(0.03)	(0.06)
Enrollment	0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Debt PPE	0.02**	0.07	0.01**	0.09+
	(0.00)	(0.06)	(0.00)	(0.04)
Teacher Salary	-0.00	-0.00	-0.00*	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
% Free Lunch	-0.00**	-0.00	-0.00**	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
% Minority	-0.00**	-0.00	-0.00	-0.00
	(0.00)	(0.01)	(0.00)	(0.00)
% Limited English Proficiency	0.00	0.01	-0.02*	-0.02*
	(0.01)	(0.01)	(0.01)	(0.01)
% Students with Disabilities	-0.83**	-1.10+	-0.98**	-1.34*
	(0.25)	(0.57)	(0.26)	(0.63)
Constant	0.30+	0.26	0.54**	-0.17
	(0.17)	(0.44)	(0.16)	(0.37)
Observations	4,489	809	4,489	809
R-squared	0.02	0.07	0.03	0.09
Number of District	651	116	651	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x
Robust standard errors in parentheses ** p<0.01, * p<0.05, + p<0.1				

Table 5: Effects of C4E on Academic Outcomes (Unstandardized)				
	(1)	(2)	(3)	(4)
VARIABLES	Math	Math	English	English
Treatment*Post	1.56**	-0.51	2.78**	0.14
	(0.51)	(0.75)	(0.48)	(0.74)
Enrollment	0.00	-0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Debt PPE	0.25**	1.18*	0.24**	1.13**
	(0.06)	(0.58)	(0.07)	(0.42)
Teacher Salary	-0.04**	-0.05	-0.08**	-0.01
	(0.01)	(0.03)	(0.01)	(0.02)
% Free Lunch	-0.01	-0.04	0.05**	-0.01
	(0.02)	(0.04)	(0.02)	(0.03)
% Minority	-0.06**	-0.06	-0.05**	-0.06
	(0.01)	(0.06)	(0.02)	(0.06)
% Limited English Proficiency	0.27*	0.36**	0.30*	0.28*
	(0.12)	(0.09)	(0.14)	(0.14)
% Students with Disabilities	-10.85**	-14.32*	-10.73**	-14.41*
	(2.92)	(6.38)	(3.00)	(6.43)
Constant	670.53**	664.52**	669.70**	652.35**
	(2.14)	(5.37)	(2.48)	(5.46)
Observations	4,489	809	4,489	809
R-squared	0.85	0.89	0.50	0.66
Number of District	651	116	651	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x
Robust standard errors in parentheses				
** p<0.01, * p<0.05, + p<0.1				

Table 6: Effects of C4E on Institutional Variables (Teacher-Student Ratio and Per Pupil Expenditures)

VARIABLES	(1) Teacher-Student Ratio	(2) Teacher-Student Ratio	(3) PPE	(4) PPE
Treatment*Post	-0.05	-0.07	-0.61*	-0.16
	(0.11)	(0.14)	(0.31)	(0.33)
Enrollment	0.00**	0.00**	-0.00**	-0.00**
	(0.00)	(0.00)	(0.00)	(0.00)
Debt PPE	-0.00	0.09		
	(0.01)	(0.07)		
Teacher Salary	0.14**	0.16**		
	(0.01)	(0.02)		
% Free Lunch	0.01**	0.01	-0.00	0.01
	(0.00)	(0.01)	(0.01)	(0.01)
% Minority	-0.01**	0.00	-0.00	-0.04+
	(0.00)	(0.01)	(0.00)	(0.03)
% Limited English Proficiency	-0.00	-0.01	-0.04	-0.06
	(0.02)	(0.04)	(0.04)	(0.05)
% Students with Disabilities	-0.02**	0.00	0.02*	-0.01
	(0.00)	(0.01)	(0.01)	(0.01)
Constant	-1.21+	-3.96*	26.91**	29.77**
	(0.64)	(1.86)	(0.82)	(1.80)
Observations	4,489	809	4,489	809
R-squared	0.72	0.70	0.39	0.32
Number of District	651	116	651	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

Table 7: Effects of C4E on Revenue Categories (State-Aid and Local Revenue)

VARIABLES	(1) Local Revenue	(2) Local Revenue	(3) State Aid	(4) State Aid
Treatment*Post	-0.90**	-0.58+	0.41**	0.44**
	(0.29)	(0.32)	(0.09)	(0.13)
Enrollment	-0.00**	-0.00**	-0.00**	-0.00**
	(0.00)	(0.00)	(0.00)	(0.00)
% Free Lunch	-0.01	0.00	-0.00	0.01
	(0.00)	(0.01)	(0.00)	(0.01)
% Minority	0.00	-0.02	-0.01*	-0.03*
	(0.00)	(0.02)	(0.00)	(0.01)
% Limited English Proficiency	-0.06+	-0.08	0.00	-0.01
	(0.03)	(0.05)	(0.02)	(0.03)
% Students with Disabilities	0.01*	0.01	-0.00	-0.02+
	(0.01)	(0.01)	(0.00)	(0.01)
Constant	14.36**	14.18**	10.02**	11.99**
	(0.66)	(1.58)	(0.30)	(0.71)
Observations	4,489	809	4,489	809
R-squared	0.23	0.11	0.42	0.61
Number of District	651	116	651	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

Appendix:

Notes on Variable Construction

The control variables include percent of students eligible for free lunch (a proxy for low-income status), percent of students from a racial/ethnic minority group, percent of students with limited English proficiency (LEP), percent of students with disabilities, student enrollment, other revenue per pupil, debt payments per pupil and average teacher salary. Five of these variables were calculated by the authors, and the rest were directly reported from New York State. Percent minority is calculated by subtracting percent white from 100. Debt payments per pupil are calculated by summing debt service interest payments and debt service principle payments and dividing by enrollment. Average teacher salary is calculated by dividing total expenditures on teacher salaries by the number of teachers. All financial variables are adjusted for inflation to year 2015 dollars. Finally, percent of students with disabilities is calculated by dividing the total number students with disabilities in grades 3-8 by total students in grades 3-8.A verage Teacher Salary approximately 15 outliers greater than $150,000, which we remove from the sample. Percent free lunch eligibility included one value greater than 100, which was replaced with the mean of the year prior and year post observations.

Placebo Tests

The placebo test was carried out by employing our difference-in-differences models during the 2005-06 and 2006-07 pre-treatment period. Our treatment indicator was changed to the binary indicator placebo, which was coded 1 if a district was in the C4E treatment group and the year was 2006-07, and 0 other. This indicator constitutes a “fake” treatment affecting C4E districts in 2006-07, a treatment that did not occur in reality. Analyzing the impact of this artificial treatment allows us to assess whether C4E districts experienced different trends in our dependent variable in the pre-treatment period. A summary of the placebo indicator is included in Appendix Table 1. Tables with the outputs of our placebo tests are included in Appendix Tables 2, 3, 4, and 5.

Appendix Table 1: Propensity Score Regression	
	(1)
	C4E
VARIABLES	Treatment
Math	-1.17
	(0.63)
English	0.46
	(0.55)
Math (Poor)	0.75
	(0.43)
English (Poor)	-0.75
	(0.34)
Student-Teacher Ratio	0.22
	(0.06)
PPE	0.02
	(0.13)
Local Revenue PP	-0.01
	(0.13)
State Aid PP	-0.24
	(0.14)
Debt PPE	0.12
	(0.15)
Teacher Salary	-0.05
	(0.01)
Enrollment	0.00
	(0.00)
% Free Lunch	0.03
	-0.16
% Minority	-0.01
	(0.01)
% Limited English Proficiency	0.08
	(0.04)
% Students with Disabilities	0.05
	(.03)
Constant	-1.59
	(0.89)
Observations	534
R-squared	0.41
standard errors in parentheses	

Appendix Table 2: Summary of Placebo Indicator

Variable	Obs	Mean	Std. Dev.	Min	Max
Placebo	1,261	0.05	0.21	0.00	1.00

Appendix Table 3: Placebo Tests for Institutional Variables (Teacher-Student Ratio and Per Pupil Expenditures)

VARIABLES	(1) Teacher-Student Ratio	(2) Teacher-Student Ratio	(3) PPE	(4) PPE
Placebo	0.17+	-0.14	0.15	0.35
	(0.09)	(0.13)	(0.54)	(0.60)
Observations	1,261	229	1,261	229
R-squared	0.91	0.92	0.18	0.08
Number of District	649	116	649	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

Appendix Table 4: Placebo Tests for Revenue Categories (State-Aid and Local Revenue)

VARIABLES	(1) Local Revenue	(2) Local Revenue	(3) State Aid	(4) State Aid
Placebo	0.21	0.49	-0.04	0.04
	(0.48)	(0.54)	(0.05)	(0.07)
Observations	1,261	229	1,261	229
R-squared	0.13	0.05	0.38	0.62
Number of District	649	116	649	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

Appendix Table 5: Placebo Tests for Academic Outcomes (Standardized)

VARIABLES	(1) Math	(2) Math	(3) English	(4) English
Placebo	0.01	0.06	-0.04	0.06
	(0.04)	(0.05)	(0.04)	(0.05)
Observations	1,261	229	1,261	229
R-squared	0.02	0.12	0.04	0.23
Number of District	649	116	649	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

Appendix Table 6: Placebo Tests for Academic Outcomes (Unstandardized)				
VARIABLES	(1) Math	(2) Math	(3) English	(4) English
Placebo	**0.66**	**0.55**	**0.94***	**0.65**
	(0.45)	**(0.64)**	**(0.46)**	**(0.63)**
Observations	1,282	231	1,282	231
R-squared	0.79	0.84	0.55	0.62
Number of District	649	116	649	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x
Robust standard errors in parentheses ** p<0.01, * p<0.05, + p<0.1				

Appendix Table 7: Performance Effects adjusted for Resources (Standardized)				
VARIABLES	(1) Math	(2) Math	(3) English	(4) English
Treatment*Post	-0.01	-0.11	-0.06	-0.14*
	(0.05)	(0.06)	(0.03)	(0.06)
PPE	0.02**	0.02*	0.01+	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Observations	4,489	809	4,489	809
R-squared	0.02	0.07	0.03	0.09
Number of District	651	116	651	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x
Robust standard errors in parentheses ** p<0.01, * p<0.05, + p<0.1				

Appendix Table 8: Heterogenous Effects by Accountability Status (Standardized)				
VARIABLES	(1) Math	(2) English	(3) Math	(4) English
Treatment*Post	-0.01	-0.06	-0.10	-0.15*
	(0.06)	(0.04)	(0.08)	(0.07)
Accountability	-0.02	-0.01	-0.01	0.02
	(0.05)	(0.04)	(0.05)	(0.04)
Observations	4,489	4,489	809	809
R-squared	0.02	0.03	0.07	0.09
Number of District	651	651	116	116
Full Sample	x	x		
Matched Sample Treatment Only			x	x
District FE	x	x	x	x
Year FE	x	x	x	x
Robust standard errors in parentheses ** p<0.01, * p<0.05, + p<0.1				

Appendix Table 9: Equity Effects (Performance Effects for Economically Disadvantaged Students)				
VARIABLES	(1) Math	(3) Math	(4) English	(6) English
Treatment*Post	-0.04	-0.15	-0.08	-0.20*
	(0.07)	(0.10)	(0.07)	(0.10)
Observations	3,929	799	3,929	799
R-squared	0.02	0.06	0.02	0.08
Number of District	609	116	609	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x
Robust standard errors in parentheses ** p<0.01, * p<0.05, + p<0.1				

Appendix Table 10: Equity Effects Placebo Test

VARIABLES	(1) Math	(2) Math	(3) English	(4) English
Placebo	0.03	0.09	-0.06	0.08
	(0.06)	(0.08)	(0.07)	(0.10)
Observations	1,078	226	1,078	226
R-squared	0.01	0.11	0.06	0.30
Number of District	577	116	577	116
Full Sample	x		x	
Matched Sample		x		x
District FE	x	x	x	x
Year FE	x	x	x	x

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1