Illusory Effects of Performance Management:

The Case of Contracts for Excellence in New York School Districts

Philip Gigliotti and Lucy C. Sorensen

University at Albany, SUNY

Abstract

Performance management systems couple outcomes-based accountability with strong managerial reforms. While a rich literature documents the behavior of managers within these systems, the literature on the performance effects of performance management is less conclusive. Since the introduction of the No Child Left Behind Act, the public education sector has heavily emphasized outcomes-based accountability systems centered on student test performance. We identify a unique performance management inspired reform that contractually required 58 New York State public school districts to develop individualized management reform plans based on "best practices" and report compliance and performance measures back to the state. We evaluate this system's impact on institutional decision-making and organizational performance using a difference-in-differences approach. We uncover negative or precisely estimated null effects on math and English test performance ranging from 0 to negative .14 standard deviations. Furthermore, we uncover evidence of multiple undesirable institutional responses that could have compromised the performance outcomes of this reform. This study contributes generalizable evidence on the way that institutional responses to performance management systems affect the implementation and ultimate success of these reforms.

Introduction

During the last quarter of the twentieth century, public sector reformers renewed their search for ways to improve the efficiency of public sector organizations. They believed that these organizations lacked the discipline imposed on private sector organizations by market forces. Since public sector organizations lacked these economic incentives, they sought to implement

other methods of incentivizing proper organizational management. This philosophy of public sector reform, known as New Public Management (NPM), was the source domain for the predominant accountability based reform of this era, performance management (Moynihan 2006). Under performance management, external stakeholders hold public managers accountable to rigorous performance standards, while also encouraging entrepreneurship by empowering managers with the discretion to lead broad organizational reforms.

Public management scholars responded to these reforms by studying organizational behavior under performance management. The main approach in this literature viewed the performance information use as integral to the success of performance management systems and sought to understand its determinants (Kroll 2015, Moynihan and Pandey 2010). Another stream of research explored "partial implementation" of performance management systems, in which managers are not provided the managerial authority necessary to drive meaningful organizational change (Moynihan 2006, Nielsen 2013). The key assumption underlying both theoretical perspectives is that effective performance information use and provision of managerial authority are essential to performance management – and that individual performance management reforms are unlikely to succeed if either fails.

During this time, 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 1970's and 1980's. The first nationwide accountability systems were initiated in the Clinton era, and culminated in the No Child Left Behind Act (NCLB) under the Bush administration (Patrick and French 2011, West and Peterson 2003). While these reforms did allow some managerial discretion at the state level

in developing accountability systems, they were not true performance management systems, in that they were not focused on providing managers with the discretion to pursue broad organizational change (Ladd 2012, 2017, Dee et al. 2010). In recent years, some local school systems, such as the New York City Department of Education (NYDOE), have implemented performance management systems with greater managerial discretion, showing qualified positive effects (Sun and Van Ryzin 2014, Wang and Yeung 2017, Destler 2016, Childress et al. 2011).

Despite an era characterized by the implementation of performance management systems in a broad range of public sector contexts, evidence of their impact on organizational performance has often been considered inconclusive due to limited study of the subject in the public management literature. Only in the past 10 years have a number of studies emerged investigating the performance impacts of performance management (Hvidman and Andersen 2013, Nielsen 2013, Poister, Pasha, and Edwards 2013, Carlson, Cowen, and Fleming 2013, Walker, Damanpour, and Devece 2011, Andersen 2008, Kroll 2016). An expansive review of the literature by Gerrish (2016) suggested small average effects of performance management systems, but found larger effects in studies adhering to methodological best practices. However, this literature remains inconclusive due to heterogeneity of results and approaches. While a number of studies do show positive effects of performance management (Poister, Pasha, and Edwards 2013, Carlson, Cowen, and Fleming 2013, Walker, Damanpour, and Devece 2011), others show null, inconclusive or qualified results (Hvidman and Andersen 2013, Andersen 2008, Kroll 2016)

This study contributes to the literature on performance effects of performance management by evaluating the case of the Contracts for Excellence (C4E) program in New York State (NYS). Under C4E, NYS invested in the managerial transformation of 58 underperforming

school districts. In exchange for enhanced accountability and oversight, C4E districts were provided 30% state aid funding increases and allowed to create individualized managerial reform plans. The state identified five evidence-based reforms (class size reduction, increased time on task, teacher and principal training, improved pre-k and kindergarten programs, and middle school and high-school restructuring), and then allowed districts to create individualized managerial reform plans constructed from this "menu" of choices. To increase managerial autonomy, districts were allowed to allocate up to 15% of C4E resources in "experimental" reforms originated by educators in their district. While they were allowed significant discretion to construct plans that reflected the unique conditions of their districts, all districts were required to include class size reduction in their plans and target resources to low-performing student groups and schools. Final plans were subject to approval from the state education department (New York State Education Department 2018).

Each year, districts were subject to oversight and random screening to ensure that resources were spent according to these plans, and at year end districts were assessed based on performance. Furthermore, continued funding was contingent of the delivery of a new plan annually, and the faithful execution of its objectives. The program also required that districts make their plans and performance metrics public and allow mechanisms to facilitate public comment. Districts making significant academic progress then had the potential to "graduate" from the program, at which point they would no longer be subject to accountability oversight, but would retain increased levels of state aid (New York State Education Department 2018).

This study describes an empirical evaluation 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, we are able to test a number of hypotheses relating to the institutional and performance effects of the reform.

We first calculate direct treatment effects on managerial decision-making and organizational performance, and then modify our model to disentangle the causal mechanisms of the program, contributing evidence on how funding, management and accountability matter to performance management effects. Using a quasi-experimental difference-in-differences approach with district and year fixed effects, we derive estimates of the effects of this program on both organizational behavior and organizational performance. We find that while this performance reform delivered additional resources to targeted districts, treated districts responded to these investments by reducing local revenue collection, leading to negative or null impacts on institutional resources. This finding may explain the lack of measured institutional response to the reform. In terms of institutional performance, we find that the reform lead to no true performance gain. However, we find that unadjusted estimates of performance improvements produce a spurious positive effect, which could have confounded accountability estimates. Through further analysis we determine that accountability failures and ineffective managerial reform, rather than resource deficits are the most likely explanation for the failure of the C4E program.

The findings in this paper contribute a number of insights to the theoretical understanding of performance management. While they should not be taken as evidence that performance management reforms don't work, they do provide evidence on ways that institutional responses can work against the desired outcomes of performance management systems. In this piece we provide direct empirical evidence of perverse institutional responses such as declining local revenue collection and adaptation to performance measurement. By providing a detailed institutional analysis that goes beyond basic performance effects, and delves deeply into institutional responses and causal mechanisms, we present compelling evidence on

implementation failures that can help to explain heterogeneous and sometimes contradictory patterns from the performance management literature.

Performance Management and K-12 Education

In the era of performance management, accountability-based reforms have grown in popularity as a solution for the perceived problems of inefficiency and lack of market discipline in public sector organizations. Beginning with the Government Performance Results Act of 1993 (GPRA), which stipulated that government agencies set measurable goals and submit annual performance reports, reformers have sought to align the incentives of public managers with those of their external stakeholders (Kravchuk and Schack 1996). However, evaluation of these early efforts quickly demonstrated that outcome-based accountability alone was insufficient to transform public sector organizations. Instead, reformers realized that accountability must be coupled with a strong management component, in which public managers used performance information to lead organizational change (Kravchuk and Schack 1996, Hatry 2002).

This insight drove significant research as scholars sought to understand how "management matters" to organizational performance. Scholars such as Moynihan (2006), Moynihan and Pandey (2005), and Meier and O'Toole (2002) built a significant corpus of research documenting how public management influenced organizational performance under performance management systems. One current within this literature sought to empirically model how managers behave under a number of environmental and organizational conditions (Meier and O'Toole 1999, Ryu 2016). Others discussed how performance information was used by public managers to drive organizational change (Heinrich 1999, 2002, Moynihan and Pandey 2010). This research contributed a nuanced understanding of how managers make decisions and use information under performance management systems.

Despite extensive investigation into management dynamics and informational systems, the public management literature was less robust in demonstrating the causal impacts of performance management systems on organizational performance. Only recently has this line of inquiry been consistently pursued. While results are somewhat inconsistent, a number of studies do indicate a positive impact of performance management. The impacts are typically heterogenous by organizational characteristics and fidelity of implementation of the reform. In the public management literature, studies by Hvidman and Andersen (2013), Andersen (2008) and Nielsen (2013) present conflicting views of the efficacy of performance management in Danish public schools, with one findings null effects in public schools and the others 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 the current state of research, inferring small positive effects of performance management systems on organizational performance – but these effects are larger 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.

Another source of evidence for the impacts of accountability-based management reforms can be found in the educational policy literature, which has documented the pervasive implementation of outcomes-based accountability reforms in public schools during the era of New Public Management. Beginning in the Clinton era, and culminating in the No Child Left

Behind Act (NCLB) under George W. Bush, public school districts nationwide implemented far-reaching accountability-based management reforms (West and Peterson 2003, Patrick and French 2011). While these reforms intended to link strong accountability measures with managerial discretion, they often fell short in practice by focusing predominantly on accountability and neglecting the management component (Ladd 2017).

Nonetheless, these new policies offered an important opportunity to evaluate the effects of outcomes-based accountability 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 impact on achievement. More commonly, scholars focused on the perverse incentives and unintended consequences resulting from the specific mechanisms and measured outcomes chosen under NCLB's accountability system (McMurrer 2007, Dee et al. 2010, Ladd and Lauen 2010). Unlike the C4E program considered in the current study, NCLB accountability standards and sanctions were not matched with promises of increased revenues or requirements for structured managerial reform (Dee and Jacob 2011).

Outcomes-based accountability and performance management systems are also found at the local level. 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 2017), found qualified evidence of impacts on student achievement (Sun and Van Ryzin 2014).

The current study offers quasi-experimental evidence from a performance management reform from New York State (NYS), an outcomes-based accountability program in the public education sector known as Contracts for Excellence (C4E). According to the performance management model, this system paired specific accountability measures with managerial reforms to steer organizational change. In exchange for increased state aid, districts created individualized management reform plans based on best-practices educational interventions and were held accountable to both the state and the public for successful execution of the plans and improved student outcomes. This is a noteworthy program, because not only did stakeholders attempt to improve organizational outcomes by incentivizing managerial reforms and setting accountability targets, but they promised financial resources to support the local districts in achieving those targets. Thus, insights presented by this case can inform our understanding of how resource adequacy matters for the efficacy of performance management systems.

Contracts for Excellence (C4E) in New York State

Education finance in New York State (NYS) has been a contentious issue for many years. In the mid 1990's, despite property tax rates that were among the highest in the nation, under-privileged schools demonstrated persistently substandard performance and experienced persistent fiscal inequality (Duncombe and Yinger 2000). In 1993, the Campaign for Fiscal Equity was formed as an advocacy group determined to address the situation in underperforming NYS schools through financial remediation. Their advocacy led to years of litigation, culminating in a landmark ruling in the case of Campaign for Fiscal Equity vs. the State of New

York (2003). The subsequent Education Budget and Reform Act of 2007 emerged, implementing a new state aid formula based on the estimated cost of closing performance gaps. In addition, they 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 (Abbott, Hodgens, and Wenzel 2013).

The new state aid formula led to dramatic increases in funding to low performing districts. 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 state aid 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 state aid increases, but their funding levels were maintained at the level of the 2008-09 academic year (New York State Education Department 2018).

In addition to providing revenues, the C4E program put a performance management system into action. Frederickson & Frederickson (2006) argue that performance management reforms often fail because public organizations are not provided with the resources to drive meaningful organizational reform. In the case of C4E, however, not only were districts required to lead management reforms to meet accountability targets, they were ostensibly given resources to do so. In exchange for state aid investments districts developed individualized management reform programs for their districts 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 managerial 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 constrained by best-practice guidelines, in practice there was wide variation in the programs that developed. (New York State Education Department 2018)

Districts were required to submit revised plans at the beginning of each academic year, and were also required 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.) Early completers were generally school districts that only struggled with

performance deficits for a small number of specific subgroups whose problems could be more easily remedied. 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 2018).

Theoretical Framework

The performance management literature demonstrates that "management matters" to public organizations, and that managers have the power to shape organizational performance (Moynihan 2006, Meier and O'Toole 2002, Moynihan and Pandey 2005). It provides evidence that managers are influenced by both internal and external stimuli and environmental factors (Meier and O'Toole 1999, Ryu 2016). Furthermore, it shows that performance management systems can influence the way that managers use information to shape managerial decisions (Heinrich 1999, 2002, Moynihan and Pandey 2010). While this body of research establishes that managers are important for organizational performance, and that managers change their behavior in response to performance management systems, there is a gap in research linking the two pieces in the causal chain to show whether performance management systems can improve managerial practice and organizational performance.

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. Gerrish (2016) suggests 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). Performance management systems have many critics, whose arguments are bolstered by the lack of consensus that these programs can meaningfully impact organizational performance. A better understanding of whether performance management impacts performance, and how controversial aspects such as accountability targets contribute to those impacts, can help public management scholars balance criticism of these systems with their overwhelming appeal to policymakers.

By 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, we have a unique opportunity to drive theory forward. For the C4E program to translate into improved downstream organizational performance, a prerequisite would be for the accountability measures to successfully drive management reforms and institutional practices. To test whether C4E had an institutional impact on NYS school districts we introduce our first hypothesis:

$\boldsymbol{H_1}$: *The C4E performance management system caused school districts to change their institutional behavior in a manner consistent with their management reform plans.*

The primary institutional 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, we can test whether C4E led to successful

managerial reform by measuring its effect on class size. Therefore, we test the hypothesis that the C4E program reduced student-teacher ratios in affected districts. We also examine the extent to which district resources and revenues change in response to the C4E program.

If C4E had a substantive institutional impact on school districts, then we would expect to see this effect carry through to measures of organizational performance. Such a result would be theoretically important. It would help clarify conflicting findings from the public management literature (Nielsen 2013, Hvidman and Andersen 2013) and help move the literature towards convergence on an understood impact of performance management. In the context of an educational organization, the most salient measure of performance is student achievement. If C4E improves organizational performance we should see measurable impact on student outcomes. This leads to our second hypothesis:

$\boldsymbol{H_2}$: *The C4E system directly improved organizational performance in the form of district-level student achievement outcomes.*

In other words, this hypothesis tests in the most direct manner possible whether the managerial incentives explored through hypothesis $\boldsymbol{H_1}$ successfully carried through to organizational performance, which we operationalize through student math and English proficiency.

The opportunity to conduct a rigorous program evaluation of the C4E program and estimate its overall impact is a significant opportunity to contribute to the literature, as there are few case studies available that have provided such an estimate. However, to be of broader interest to the public management literature, we seek to disentangle the main causal mechanisms of the program to understand how money, management and accountability matter to performance management systems (See Figure 1). In particular, we investigate three mechanisms, which are

not mutually exclusive, through which C4E shapes organizational practices and performance, specified below:

$\boldsymbol{H_{3A}}$: *The effects of the C4E program are attributable to the large 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. We must test, therefore, whether increased educational expenditures and reduced class size, which have been shown to improve student achievement (Jackson, Johnson, and Persico 2015, LaFortune, Rothstein, and Schanzenbach 2016) drive program effectiveness ($\boldsymbol{H_{3A}}$**).** A second major component of the C4E program, which is common among performance management systems, is outcomes-based accountability measurement ($\boldsymbol{H_{3B}}$**)**. 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, we introduce a series of modifications to our model (described fully in the results section) to parse out the different factors at play. To test the impact of financial resources on C4E district performance, we estimate our models conditional on the level of per-pupil expenditures (PPE). To investigate the role of accountability

requirements, we exploit a feature of the program in which graduated districts are no longer subjected to heightened accountability measurement and reporting, to observe whether performance trends shifted after "graduating" C4E. Finally, we infer that residual effects not attributable to either increased financial resources or to heightened accountability reflect pure management effects. Figure 1 illustrates each of these potential pathways through which C4E could improve district performance and/or create more equitable outcomes.

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. We wonder whether this equity objective plays out as intended at the district managerial level: Do traditionally underperforming and disadvantaged student groups improve more than their peers in response to the C4E system? This leads to our hypothesis:

$\boldsymbol{H_4}$: *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.*

Educational policy analysts often describe educational equity through the construct of vertical equity, which acknowledges that different groups of students have different characteristics and needs, and prescribes differential treatment to equalize outcomes between them (Berne and Stiefel 1984). 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). Testing the distributional impacts of C4E allows us to assess whether the equity components of the program "had teeth," or whether districts could circumvent them, relying

mainly on demonstrating improvement through performance gains among non-disadvantaged students. To test this hypothesis, we exploit rich performance data that includes district-level average test scores restricted to disadvantaged student populations targeted under C4E (economically disadvantaged, students with disabilities, limited English proficiency students).

The hypotheses motivating this study contribute to central questions in the performance management literature that have not received sufficient empirical attention. We expand upon prior work by assessing – at scale – a performance management reform with a number of unique characteristics that can inform future practice. We extend the fundamental question of whether performance management systems work to disentangle the causal mechanisms underlying observed effects, and to further consider whether equity in performance can be effectively targeted. As described below, we carefully evaluate each hypothesis using quasi-experimental methods and rich panel data.

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 Data Site, specifically the School Report Cards, which contain both school- and district-level data. Merging these data sources together results in a complete match to the 677 major school districts included in the FARU data. Because some of these districts only serve elementary or high school students, our 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. (The data on NYC schools was

furthermore incompatible for matching to the fiscal database.) 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 institutional resources and resource allocation. We use student-teacher ratio (student enrollment divided by the total number of teachers) and per-pupil expenditures (PPE) (total expenditures divided by enrollment) to measure institutional change resulting from the program. We also examine local revenue per-pupil (total local revenue divided by enrollment) and state-aid per-pupil (total state-aid revenue divided by enrollment).

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. Our measure consists of averaged performance of students in grades 3-8. These scores are scaled from 0 to 800, but in practice most scores are clustered between 600 and 700. We also make use of 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. For each test performance measure we consider both raw scores and scores standardized to have mean zero and standard deviation of one by school year. It is important to test both raw and standardized scores to ensure that any observed treatment effects are not being driven by distributional changes in the dependent variable, which could lead to spurious results.

This study uses a binary treatment indicator of whether a school district was a C4E district during any observed year ($Treatment$). This variable equals one if the district was ever a part of Contracts for Excellence and zero if it was not. The treatment group consists of 58 school districts that participated in C4E over the course of the program. Our model also includes a binary policy indicator that equals one if the district participated in C4E and the time period is post 2007-08 school year implementation ($Treatment * Post$). Though some districts departed the original C4E cohort as the program developed, at which point they were no longer subject to accountability requirements, we do not remove districts from the policy treatment group as they leave in our 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. We provide empirical evidence for these assumptions in the analytical section of this text. In addition, our sample includes three districts that entered the program one year late in 2008-09. We modify our $Treatment * Post$ indicator to be coded zero during the 2007-2008 academic year for these 3 districts.

Finally, the dataset includes a robust 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 provided 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 effect of Contracts for Excellence on 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 difference-in-differences model according to the following model:

$$y_{dt} = \beta_0 + \beta_1 Post_t + \beta_2 Treatment_d + \beta_3 Post \times Treatment_{dt} + \beta_4 X_{dt} + \varepsilon_{dt}$$ (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 treatment indicator accounts for all time-invariant characteristics of C4E districts that distinguish them from other districts, and including the post indicator accounts for all

unobserved characteristics of the post-implementation period. This functional form mitigates the selection problem of the naïve model and allows β_3 – the coefficient of interest – to capture effects of the C4E reform with greatly reduced bias.

Equation 1 refers to the classic difference-in-differences approach which assumes one uniform pretreatment period, and one uniform post-treatment period. However, this model will yield noisy and potentially biased estimates in a panel data structure with multiple years in both pre- and post-treatment periods. This assumption can be relaxed by simply adding a full vector of year fixed effects τ_t and removing the post indicator from the model (Wooldridge 2007). The year fixed effects control for all observed and unobserved changes across time in our sample. We can also substitute the treatment indicator with θ_d, a vector of district fixed effects. Adding district fixed effects controls for all observed and unobserved time-invariant characteristics of the district, including the fact that it is a C4E district ($Treatment_{dt}$), which we therefore remove from the model. A generalized form of the difference-in-differences model with multiple periods can therefore be specified as a two-way fixed effects model according to the following equation:

$$y_{dt} = \gamma_0 Post \times Treatment_{dt} + \gamma_1 X_{dy} + \theta_d + \tau_t + \varepsilon_{dt} \quad \text{(Equation 2)}$$

The estimates generated by this model most completely account for selection bias into the C4E program and for any contemporaneous events. This is our preferred model.

Each model is 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, our empirical strategy relies on the parallel trends assumption: that C4E district pre-treatment trends were similar to those of non-C4E districts. We conduct a series of placebo tests during the pre-treatment period in order to investigate this assumption, with details provided in the results section.

Another potential concern is that C4E districts differ systematically from other districts in the state in ways that could drive differential trends in student performance in the post-reform period. For this reason, we estimate the same model described above but 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, we estimated propensity scores using a probit regression restricted to the last pre-treatment year (2007) and including all available covariates listed in Table 1. We then matched each C4E district to its nearest neighbor without replacement, and absorbed the entire panel of each match to form a comparison group. This allows us to estimate treatment effects compared to a similar comparison group, against whom we should expect to see a genuine effect. In our results in the following section, we refer to our models estimated against the full sample as Model 1 and those estimated against the matched sample as Model 2.

Results

The objective of this analysis is to first estimate the effects of C4E on institutional responses to determine if C4E had the intended effect on operations within school districts, namely reductions in class sizes and increases in per-pupil expenditures (PPE), and then estimate effects on multiple measures of student performance. For each outcome we estimate a generalized two-way fixed effects model as specified in Equation 2. Because C4E districts were not selected randomly, but rather chosen for their particularly low baseline performance, we estimate treatment effects compared to two different counterfactual comparison groups; the first is the full sample of NYS school districts (Model 1), and the second is a sample of "comparable" districts chosen through propensity-score matching at baseline (Model 2). Because a genuine

treatment effect would likely lead to a change in performance relative to similar school districts, we are most interested in the estimate compared against the matched district comparison group.

Effects on Institutional Characteristics

The two institutional 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 an indication of the implementation of the targeted managerial reform. Since the program promised increased revenues to treated districts, the PPE measure provides an indication of whether these revenues translated into increased institutional resources. Table 3 provides the results of these analysis. In both Model 1 and Model 2, C4E districts did not decrease their student-teacher ratios. This is a troubling result that indicates the managerial reforms under C4E may not have had the teeth to effect desired organizational change. We next examine treatment effects on educational expenditures (PPE). If the C4E program did not lead to increased institutional resources, this could explain the previously observed lack of expected managerial responses. The results of these analyses, also in Table 3, are again concerning. C4E treatment districts actually demonstrated lower PPE by approximately $600 in the Model 1, a relationship that is statistically significant at the .05 level. In Model 2 the estimate is a decrease of approximately $300, 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 surprising.

To better understand this finding, we estimate treatment effects 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 4) are alarming. 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. The result is statistically significant at the .01 level in Model 1 and at the .05 level 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 significant literature that documents school districts responding to intergovernmental grants by decreasing local revenue (Cascio, Gordon, and Reber 2013, Gordon 2004). This finding provides a convincing explanation for the lack of institutional response under the C4E program.

Effects on Academic Outcomes

To assess the impact of the C4E reform on organizational performance, we estimate our main model on average scores in grade 3-8 math and English end-of-year exams. The results of these analyses are presented in Table 5. The results show statistically significant impacts of 1.55 points in math and 2.78 points in reading in Model 1, but null impacts in Model 2 where we would expect to see a genuine treatment effect. The fact that the results are observed in the full sample (Model 1), which compares C4E districts against both high-performing and low-performing districts, and not in the matched sample (Model 2) which only includes low-performing districts, suggests that the results are being driven by gains compared to high-performers. In our robustness checks section, we will discuss tests which show large trend violations that are worse in the full sample and statistically significant in English. This indicates that these significant positive effects in the full sample could be a false positive, driven by

secular trend violations in our performance measures that are most pronounced relative to high-performing districts.

The finding of a false positive, coupled with null effects compared to similar districts, would be extremely informative to understandings of performance management theory. If C4E districts were able to demonstrate improved performance metrics without genuine performance gains relative to similar districts, this would have confounded the accountability system underlying the C4E program and could explain a failure to produce performance improvements. To explore this finding further, we examine the overall shifts in the distribution of performance measures in Figures 2 and 3. These figures plot histograms of the distribution of 8th grade math and ELA test scores in 2006 prior to C4E and in 2013, several years after C4E implementation. These figures show a radical upward shift in test scores of initially low-performing districts during the same time that C4E reforms took place. The minimum math score increased by more than 20 points in this period (13 in English) while the maximum score in math increased by only 6 points and actually decreased in English (Appendix Table 6). Based on this analysis, it is clear that the false positive performance increases relative to high-performing districts were driven by these secular test score gains in the bottom of the distribution. Low performing districts in New York all made significant gains, which is why we see C4E districts gaining relative to the full sample, but not compared to a matched sample which includes similar districts originally clustered at the bottom of the distribution.

If New York's educational policy had truly led to 20 point gains in the bottom of the achievement distribution, such a result would have attracted nationwide attention. However, the evidence suggests that these gains were illusory. 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 gains 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). 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 we consulted trends in the NAEP over the period of our study, there was no indication of an upward trend in NYS in any performance quartile (NAEP 2018). This suggests that the performance measurement gains experienced by NYS schools during this period were not a result of true performance gains, but rather from unintended responses of the schools to the measures themselves. This is consistent with warnings from Hood (2006, 2012), who argues that performance measures incentivize organizations to "hit the target but miss the point," distorting their behavior to meet performance targets without making constructive institutional changes.

One option to correct for these secular trends in our performance measures is to perform a transformation. One common method to detrend test scores, or smooth out differences between years or grades, is to standardize data within years, forcing each year's scores to have a mean of zero and standard deviation of 1. This is a common transformation in the education policy literature, and has been used on this data in another article on New York education finance. (Gigliotti and Sorensen 2018)

We estimate all further analyses of effects of the C4E managerial reform on student performance using these transformed dependent variables. In Table 6 we estimate the same models as Table 5, with these transformed measures. These results reveal a precisely-estimated

zero effect in 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. We see negative impacts of approximately .06 SDs in the Model 1 (statistically significant at the .10 level) and a large negative effect of .16 SDs in Model 2, which is significant at the .05 level. This suggests that the C4E program actually had either a null or negative impact on organizational performance.

Further Analyses

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

In Appendix Table 7 we estimate models that control for PPE. Because C4E led to lower resources in treated districts it's 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 large decreases in resources. After controlling for spending we conclude our findings do not result from spending effects. Estimates are essentially unchanged in terms of effect size or significance. This suggests that the lack of results was not a result of the resource deficits triggered by the program, but rather from failure of the accountability measures or lack of effective managerial reform.

To check for potential accountability impacts, we estimate models that leverage variation among treated districts in length of exposure to oversight under C4E. Because some districts graduated from the program, we can compare performance in treated districts both during and after oversight, when we assume any managerial reform would have persisted but accountability would have ended. To do so we introduce an indicator of exposure to oversight that is coded one if a treatment district was currently subject to oversight and zero otherwise. When added to our main model, this variable creates a "triple-differences" approach that estimates heterogenous impacts by oversight status. The results of these analyses are included in Appendix Table 8. The results confirm no evidence of heterogenous impact by accountability status, with precisely estimated null coefficients on the accountability indicator for both test score measures. This suggests that oversight under C4E had no performance impact, consistent with the hypothesis that secular trends in the test score measures confounded accountability mechanisms. The "triple-differences" model is not perfect, because the accountability indicator could be picking up unobserved differences between program graduators and program persisters, such as lower overall performance. To add robustness, we estimate an alternate specification that restricts the

sample to only treated districts. In this model, the C4E treatment indicator drops out, and we instead compare districts to themselves before, prior and following graduation. These results also reveal zero or negative estimates, indicating oversight did not contribute a meaningful positive impact to treated districts.

The results of these analyses show that neither decreased resources nor accountability mechanisms led to substantive changes in performance. If we believe the negative treatment effects observed in our main analyses, then we are left to attribute these negative impacts to effects of the managerial interventions put in place under C4E. If the managerial reforms under the program were misguided or poorly executed or diverted resources from core mission objectives, this could explain the negative impacts observed under the program. Pandey, Pandey, and Miller (2017) recently explored managerial innovation in public school districts and found negative impacts on performance, a finding that they claimed was supported by other negative or null impacts in the literature. While innovation in schools inherently presents an avenue for improved performance, its possible that there is potential for both negative or positive impacts to develop, depending on implementation.

Our final hypothesis concerns equity effects of the C4E program. We want to see whether the possibly negative impacts observed in our 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, we estimate models with test score data from NYS 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 than in the main analyses, with very large negative effects measured in English in Model 2 sample, 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 main model, and perhaps driving them in English. This raises significant equity concerns about the possibility that performance management reforms having disproportionate 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 common referred to as parallel trends. To test this assumption, we can perform a "placebo test" in which we re-specify the empirical model using only the two pre-treatment periods and assign an "artificial" treatment to C4E districts in the year before implementation occurs. 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 us to determine whether C4E districts displayed differential trends in the dependent variables during the pre-treatment period (which could bias the main results). The results of these analyses are included in Appendix Tables 2 through 5 and 10.

First, we consider placebo tests for the models in Table 3 which show effects on student-teacher ratios and PPE. For student-teacher ratio (Appendix Table 2), the placebo coefficient for the Model 1 is marginally significant at the .10 level, while the estimate for Model 2 is statistically indistinguishable from zero. These tests lead us to place greater confidence in estimates from Model 2, which clearly indicate that the C4E program did not lead to true managerial change. For PPE, the placebo tests in Appendix Table 2 show no evidence of parallel trends violations, allowing us to interpret this result of decreased resources in both models with

confidence. Likewise, for local revenue and state-aid revenue, the placebo tests in Appendix Table 3 show no evidence of a parallel trends violation.

Next, we consider placebo tests for the academic outcomes (Appendix Table 4). As mentioned earlier in the results section, the tests on the raw test score measures in Model 1 show troubling results, with a large placebo coefficient in Math and a large significant coefficient in English. However, the coefficients in Model 2 are smaller, less precisely estimated and far from statistically significant. This suggests that the trend violations are occurring with respect to the high-performing districts that are included in the full sample, not with respect to the matched sample of districts. This is the violation responsible for the "false positive" results we observe in the full sample underlying our argument about accountability failures. After correcting the trend violation with our standardized test score measure, the placebo estimates in Appendix table 5 are all precisely estimated null coefficients in math and English in both models, although the sample size for the matched sample may not be large enough to detect a violation. We believe one should exercise some caution, therefore, in inferring a negative impact of C4E on performance. Nonetheless, we can confidently conclude from these analyses that C4E districts did not experience performance gains under the program. Given the earlier findings that C4E districts cut their local revenue effort and did not reduce average class sizes – the primary intended managerial change of the program – this absence of performance gains makes sense.

Finally, we discuss placebo tests on our test score measures which are restricted to economically disadvantaged students, which we test in Appendix Table 9. The placebo tests are included in Appendix Table 10. The placebo estimates do not raise significant concerns, but the coefficient estimates are non-zero and larger than in our main models, especially in English,

suggesting caution in inferring a negative impact from our main results as it could be driven by non-zero trend violations among economically disadvantaged students.

Conclusion

The results of our analysis contribute a worrying narrative of the way that institutional responses to performance management regimes can thwart the achievement of true performance gains. We present 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 generate measurable changes in targeted institutional outcomes. Our analysis of the performance effects of the C4E program provide 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 which tricked accountability mechanisms. After correcting our analysis for this score inflation, we estimate null or negative effects of the C4E program on math and English performance, ranging from 0 to negative .14 standard deviations, statistically significant at the .10 level in three out of four models, and at the .05 level in one out of four.

Moynihan and Pandey (2005) investigated the key drivers of performance under performance management and found that organizational buy-in, which they refer to as "developmental culture," was an important mediator for how performance management systems enhance performance. This suggests that in the context of a lack of organizational buy-in, performance management reforms would be likely to fail. The findings in our study of C4E districts reducing local revenue effort and potentially gaming test score measures suggest that

these districts may not have made a good-faith effort to enact targeted managerial reforms and achieve true performance gains under the objectives of the program. This lack of buy-in is a plausible explanation of our rejection of hypotheses H1 and H2: that districts would respond to the reform by enacting institutional change and improving performance of their schools.

To go beyond simple performance effects, we perform supplemental analyses to parse out the causal mechanisms operating in the C4E program and explain how money, accountability and management matter to this performance management system. After controlling for the perverse revenue substitution triggered by this program, we find that resource deficits did not mask performance effects of the C4E program, but rather the lack of performance impact stems from the ineffectiveness of the program itself. By exploiting variation in exposure to oversight under the program, we conclude that the accountability mechanisms under C4E were ineffectual and did not contribute to managerial enhancement. Thus, we conclude that any residual effects of the program, such as the suggestive negative impacts on math and English performance measures, were related to perverse impacts of managerial change under the program, possibly due to disruption effects resulting from poorly-resourced or implemented managerial reforms.

While we test a number of hypotheses regarding performance management, the generalizable findings in this study come from the context surrounding the results of those tests rather than the results themselves. Our results should not be interpreted as a finding that performance management systems in general do not produce performance gains. Instead, the lack of performance effects uncovered in this study suggests a number of conclusions about the causes of performance management system failure. First, providing resource incentives in the context of performance management systems can fail, as it creates a moral hazard for public organizations to decrease effort in independent revenue collection. Second, we 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 performance gains. We provide empirical evidence that the C4E program occurred in the context of potential manipulation and rapid inflation of performance measures that allowed districts to meet stringent accountability requirements.

Beyond theoretical contributions, this study makes a significant 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 even with use of robust econometric methods and rigorous testing of assumptions. For this reason, we suggest that all performance results should be compared to equivalent tests with outcome measures that have been detrended of year-to-year distributional change, by employing standardization. This method is already standard in the economics of education literature, where researchers run into these issues more frequently. The lack of attention to performance trends in dependent variables suggests that many results which use raw performance measures, or which use full-sample instead of by-year standardizations, may be unreliable.

Taken in sum, the experience of NYS school districts under C4E confirms the long-understood importance of incentives within 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 true performance improvements. This was the case under C4E, where a highly-publicized and expensive performance-based reform failed to meet its goals. 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. These results should caution both policy-makers and scholars of organizational performance to critically reflect on the role of institutional responses when designing and evaluating performance management systems.

References

Abbott, Doug, Patrick Hodgens, and Kevin Wenzel. 2013. Memorandum on New York State Education Aid Formula Reform. Center for Policy Research, Syracuse University.

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

Berne, Robert, and Leanna Stiefel. 1984. *The Measurement of Equity in School Finance*. Baltimore, MD: Johns Hopkins 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.

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., 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.

Duncombe, William, and John Yinger. 2000. "Financing higher student performance standards: the case of New York State." *Economics of Education Review* 19 (4):363-386.

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.

Gigliotti, Philip, and Lucy Sorensen. 2018. "Educational Resources and Student Achievement: Evidence from the Save Harmless Provision in New York State"." *Economics of Education Review* Forthcoming.

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.

Hatry, Harry P. 2002. "Performance measurement: Fashions and Fallacies." *Public Performance & Management Review* 25 (4):352-358.

Heinrich, Carolyn J. 1999. "Do government bureaucrats make effective use of performance management information?" *Journal of Public Administration Research and Theory* 9 (3):363-393.

Heinrich, Carolyn J. 2002. "Outcomes-based performance management in the public sector: Implications for government accountability and effectiveness." *Public Administration Review* 62 (712-725).

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).

Jackson, Kirabo C., Rucker C. Johnson, and Claudia Persico. 2015. "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms." *Quarterly Journal of Economics* 131 (1):157-218.

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. 2012. "Education and poverty: Confronting the evidence." *Journal of Policy Analysis and Management* 31 (203-227).

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.

LaFortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2016. "Title." NBER Working Paper.

McMurrer, Jennifer. 2007. Choices, changes, and challenges: Curriculum and instruction in the NCLB era. In *From the capital to the classroom: Year 5 of the No Child Left Behind Act*. Washington, D.C.: Center on Education Policy.

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

Meier, Kenneth J., and Laurence O'Toole. 1999. "Modeling the impact of public management: Implications of structural context." *Journal of Public Administration Research and Theory* 9 (4):505-526.

Meier, Kenneth J., and Laurence J. Jr. O'Toole. 2002. "Public management and organizational performance: The effect of managerial quality." *Journal of Policy Analysis and Management* 21 (4):629-643.

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., 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.

Moynihan, Donald P., and Sanjay K. Pandey. 2005. "Testing how Management Matters in an Era of Government by Performance Management." *Journal of Public Administration Research and Theory* 15 (3):421-439.

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. 2018. "Contracts for Excellence." http://www.p12.nysed.gov/mgtserv/C4E/.

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

Pandey, Sheela, Sanjay K. Pandey, and Larry Miller. 2017. "Measuring Innovativeness of Public Organizations Using Natural Language Processing Techniques in Computer-Aided Textual Analysis." *International Public Management Journal* 20 (1):78-107.

Patrick, Barbara A., and P. Edward French. 2011. "Assessing new public management's focus on performance measurement in the public sector: A look at No Child Left Behind." *Public Performance & Management Review* 35 (2):340-369.

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.

Ryu, Sangyub. 2016. "Modeling public management: Current and future research." *Public Organization Review* 16 (1):77-94.

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, Wiejie, and Ryan Yeung. 2017. "Testing the effectiveness of "Managing for Results:" Evidence from a natural experiment." Association of Education Finance and Policy, Washington D.C., March 15- March 18.

West, Martin R., and Paul E. Peterson. 2003. "The politics and practice of accountability." In *No Child Left Behind? The politics and practice of school accountability*, edited by Paul E. Peterson and Martin R. West. Washington, D.C.: Brookings Institution Press.

Wooldridge, Jeff. 2007. What's new in econometrics? Lecture 10 Difference-in-Differences estimation. edited by 2007 NBER Summer Institute: NBER Summer Institute, 2007.

Tables and Figures

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

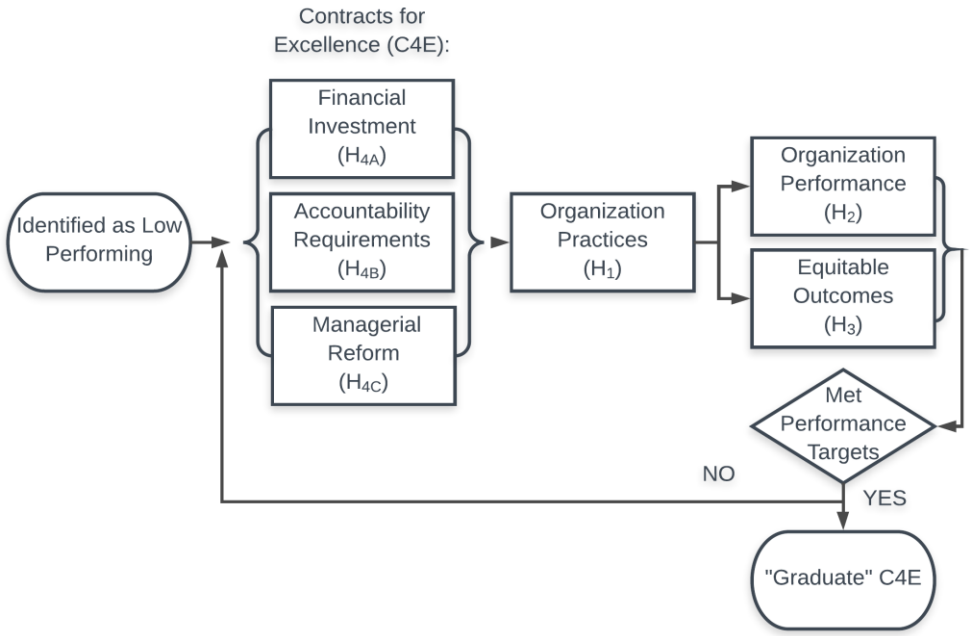

Figure 2. Histogram of District Average 8th Grade Raw Math Scores: 2006 and 2013 School Years

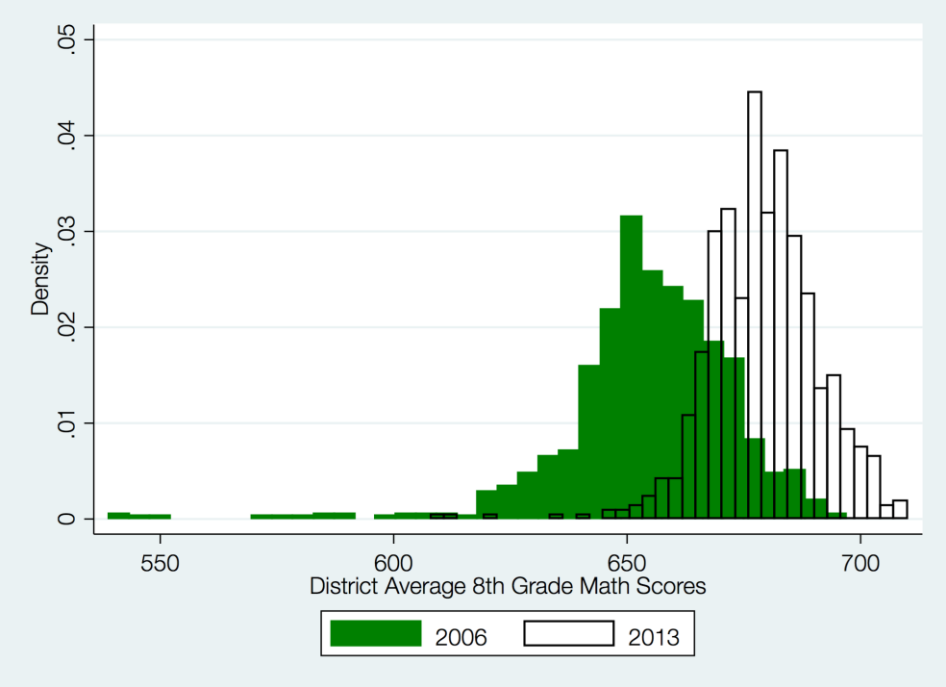

Figure 3. Histogram of District Average 8th Grade Raw ELA Scores: 2006 and 2013 School Years

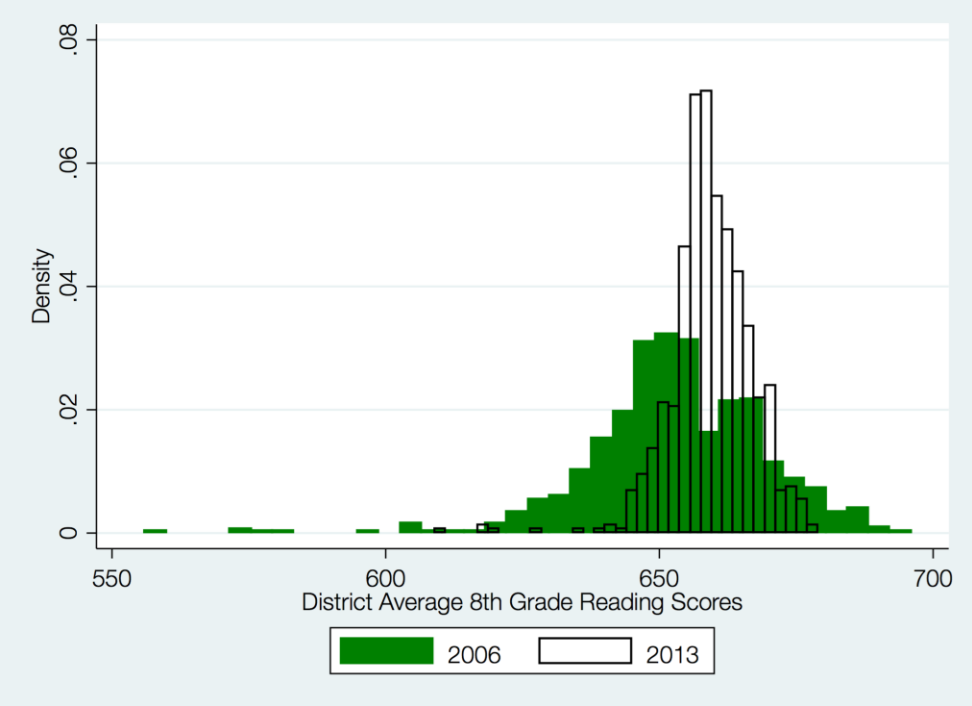

Table 1: Descriptive Statistics					
Variable	Obs	Mean	Std. Dev.	Min	Max

Math	4,512	679.84	12.22	626.50	717.00
English	4,512	667.72	8.71	630.33	696.33
Math (Standardized)	4,512	0.00	1.00	-3.58	3.41
English (Standardized)	4,512	0.00	1.00	-3.46	3.37
Math (Economically Disadvantaged, Standardized)	3,939	0.00	1.00	-4.21	4.41
English (Economically Disadvantaged, Standardized)	3,939	0.00	1.00	-3.94	4.43
C4E Treatment	4,512	0.09	0.29	0.00	1.00
Treatment*Post	4,512	0.06	0.24	0.00	1.00
Teacher-Student Ratio	4,512	11.56	1.89	4.21	30.77
PPE	4,512	20.95	4.88	12.16	83.06
Local Revenue PP	4,512	10.67	6.96	0.90	73.80
State Aid PP	4,512	8.33	3.90	0.93	23.46
Debt PPE	4,512	1.63	1.23	0.00	47.09
Teacher Salary	4,512	77.00	20.15	38.08	147.28
Enrollment	4,512	2646.44	3342.70	99.00	43436.00
% Free Lunch	4,512	22.58	14.75	0.00	94.00
% Minority	4,512	15.45	19.72	0.00	100.00
% LEP	4,512	1.91	3.98	0.00	33.00
% SWD	4,512	15.29	4.21	0.51	56.28

Table 2: Comparison between C4E and Non-C4E Districts

Factor	Non-C4E	C4E	p-value
N	1236	116	
Math	670.82 (11.70)	660.31 (11.26)	<0.001
English	666.07 (10.50)	655.94 (9.52)	<0.001
Math (Standardized)	0.08 (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	12.13 (3.45)	13.04 (1.60)	0.005
PPE	20.21 (6.55)	18.14 (4.88)	<0.001
Local Revenue PP	10.85 (8.08)	7.82 (5.75)	<0.001

State Aid PP	7.47 (3.61)	8.15 (2.80)	0.049
Debt PPE	1.40 (0.80)	1.11 (0.64)	<0.001
Teacher Salary	78.32 (19.72)	78.81 (17.94)	0.80
Enrollment	3080.39 (28999.73)	6708.86 (7836.46)	0.18
% Free Lunch	18.36 (12.83)	34.11 (16.72)	<0.001
% Minority	12.98 (18.03)	32.65 (28.32)	<0.001
% LEP	1.68 (3.42)	5.28 (7.67)	<0.001
% SWD	15.28 (6.76)	17.88 (6.22)	<0.001

Table 3: 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	**-0.06**	**-0.07**	**-0.61***	**-0.30**
	(0.11)	**(0.15)**	**(0.31)**	**(0.35)**
Debt PPE	0.00**	0.00**	-0.00**	-0.00**
	(0.00)	(0.00)	(0.00)	(0.00)
Teacher Salary	-0.00	0.00		
	(0.01)	(0.10)		
Enrollment	0.14**	0.16**		
	(0.01)	(0.02)		
% Free Lunch	0.01**	-0.00	-0.00	0.02
	(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	0.00	-0.04	-0.07
	(0.02)	(0.04)	(0.04)	(0.06)
% Students with Disabilities	-0.02**	-0.01	0.02*	0.01
	(0.00)	(0.01)	(0.01)	(0.02)
Constant	-1.22+	-3.07	26.95**	28.98**
	(0.63)	(1.89)	(0.82)	(1.80)
Observations	4,512	811	4,512	811
R-squared	0.72	0.68	0.39	0.32
Number of District	651	116	651	116
Model 1	x		x	

Model 2		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 4: 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****	**-0.65***	**0.42****	**0.44****
	(0.29)	**(0.33)**	**(0.09)**	**(0.12)**
Enrollment	-0.00**	-0.00**	-0.00**	-0.00**
	(0.00)	(0.00)	(0.00)	(0.00)
% Free Lunch	-0.00	0.01	-0.00	0.01+
	(0.00)	(0.01)	(0.00)	(0.01)
% Minority	0.00	-0.02	-0.01*	-0.02*
	(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	-0.01
	(0.01)	(0.01)	(0.00)	(0.01)
Constant	14.36**	13.66**	10.04**	12.05**
	(0.66)	(1.56)	(0.30)	(0.67)
Observations	4,512	811	4,512	811
R-squared	0.23	0.11	0.42	0.60
Number of District	651	116	651	116
Model 1	x		x	
Model 2		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

VARIABLES	(1) Math	(2) Math	(3) English	(4) English

Treatment*Post	**1.55****	**-0.78**	**2.78****	**-0.17**
	(0.51)	**(0.74)**	**(0.48)**	**(0.76)**
Debt PPE	0.00	-0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Teacher Salary	0.24**	1.04*	0.24**	0.92*
	(0.06)	(0.52)	(0.07)	(0.40)
Enrollment	-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*	0.23+	0.30*	0.20
	(0.11)	(0.12)	(0.14)	(0.14)
% Students with Disabilities	-0.11**	-0.13*	-0.11**	-0.15*
	(0.03)	(0.06)	(0.03)	(0.07)
Constant	670.67**	667.06**	669.80**	654.73**
	(2.12)	(5.10)	(2.49)	(5.21)
Observations	4,512	811	4,512	811
R-squared	0.85	0.89	0.50	0.66
Number of District	651	116	651	116
Model 1	x		x	
Model 2		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 Academic Outcomes (Standardized)

VARIABLES	(1) Math	(2) Math	(3) English	(4) English
Treatment*Post	**-0.02**	**-0.12+**	**-0.06+**	**-0.16***
	(0.05)	**(0.06)**	**(0.03)**	**(0.06)**
Debt PPE	0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Teacher Salary	0.02**	0.07	0.01**	0.08+
	(0.00)	(0.05)	(0.00)	(0.04)
Enrollment	-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)	(0.01)
% Limited English Proficiency	0.00	-0.00	-0.02*	-0.03**
	(0.01)	(0.01)	(0.01)	(0.01)
% Students with Disabilities	-0.01**	-0.01*	-0.01**	-0.02*
	(0.00)	(0.01)	(0.00)	(0.01)
Constant	0.30+	0.36	0.53**	-0.21
	(0.17)	(0.42)	(0.16)	(0.36)
Observations	4,512	811	4,512	811
R-squared	0.02	0.07	0.03	0.10
Number of District	651	116	651	116
Model 1	x		x	
Model 2		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				

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. (There were approximately 30 missing observations for the students with disabilities variable. To preserve sample integrity we used a regression based imputation using all covariates in Table 1. Results are fundamentally equivalent to models that have the original variable and 30 missing observations, but the models with the imputed variable are more complete.) Average Teacher Salary had 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: Summary of Placebo Indicator

Variable	Obs	Mean	Std. Dev.	Min	Max
Placebo	1,282	0.05	0.21	0.00	1.00

Appendix Table 2: 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+**	**-0.05**	**0.15**	**0.36**
	(0.09)	**(0.14)**	**(0.54)**	**(0.59)**
Observations	1,282	231	1,282	231
R-squared	0.91	0.90	0.18	0.08
Number of District	649	116	649	116
Model 1	x		x	
Model 2		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 3: 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**	**0.44**	**-0.03**	**0.02**
	(0.48)	**(0.53)**	**(0.05)**	**(0.07)**
Observations	1,282	231	1,282	231
R-squared	0.13	0.05	0.37	0.61

Number of District	649	116	649	116
Model 1	x		x	
Model 2		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 Academic Outcomes

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
Model 1	x		x	
Model 2		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**	**0.04**	**-0.04**	**0.04**
	(0.04)	**(0.06)**	**(0.04)**	**(0.05)**
Observations	1,282	231	1,282	231
R-squared	0.02	0.08	0.04	0.21
Number of District	649	116	649	116
Model 1	x		x	
Model 2		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: Test-Score Trends

Variable	Obs	Mean	Std. Dev.	Min	Max
Full Sample					
Math (Pre-2008)	1,282	669.87	12.04	626.50	711.17
Math (Post-2008)	3,230	683.79	9.79	647.33	717.00
English (Pre-2008)	1,282	665.15	10.81	630.33	696.33
English (Post-2008)	3,230	668.73	7.48	643.17	695.33
Low-Performers					
Math (Pre-2008)	685	661.52	7.77	626.50	678.17
Math (Post-2008)	1,734	676.76	5.95	647.33	691.33
English (Pre-2008)	685	657.41	6.50	630.33	672.00
English (Post-2008)	1,734	663.65	4.46	643.17	675.17
High-Performers					
Math (Pre-2008)	597	679.45	8.38	658.17	711.17
Math (Post-2008)	1,496	691.95	6.49	676.50	717.00
English (Pre-2008)	597	674.03	7.41	658.83	696.33
English (Post-2008)	1,496	674.62	5.76	662.00	695.33

Appendix Table 7: Performance Effects adjusted for Resources (Standardized)

VARIABLES	(1) Math	(2) Math	(3) English	(4) English
Treatment*Post	**-0.01**	**-0.12+**	**-0.06+**	**-0.16***
	(0.05)	**(0.06)**	**(0.03)**	**(0.06)**
PPE	0.02**	0.01+	0.01*	0.01
	(0.01)	(0.01)	(0.01)	(0.00)
Observations	4,512	811	4,512	811
R-squared	0.03	0.08	0.03	0.10
Number of District	651	116	651	116
Model 1	x		x	
Model 2		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	(5) Math	(6) English
Treatment*Post	-0.01	-0.06	-0.12	-0.16*		
	(0.06)	(0.04)	(0.08)	(0.07)		
Accountability	**-0.02**	**-0.01**	**-0.00**	**0.00**	**-0.03**	**-0.05**
	(0.05)	**(0.04)**	**(0.05)**	**(0.04)**	**(0.06)**	**(0.04)**
Observations	4,512	4,512	811	811	406	406
R-squared	0.02	0.03	0.07	0.10	0.11	0.08
Number of District	651	651	116	116	58	58
Model 1	x	x				
Model 2			x	x		
Treatment Only					x	x
District FE	x	x	x	x	x	x
Year FE	x	x	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	**-0.05**	**-0.17**	**-0.08**	**-0.22***
	(0.07)	**(0.10)**	**(0.07)**	**(0.10)**
Observations	3,939	801	3,939	801
R-squared	0.02	0.06	0.02	0.08
Number of District	609	116	609	116
Model 1	x		x	
Model 2		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**	**0.03**	**-0.05**	**0.05**
	(0.06)	**(0.09)**	**(0.07)**	**(0.10)**
Observations	1,087	228	1,087	228
R-squared	0.01	0.06	0.06	0.29
Number of District	577	116	577	116
Model 1	x		x	
Model 2		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				