Dissertation

Three Essays on Education Reform in the United States

Ethan Scherer

RAND CORPORATION

Dissertation

Three Essays on Education Reform in the United States

Ethan Scherer

This document was submitted as a dissertation in July 2014 in partial fulfillment of the requirements of the doctoral degree in public policy analysis at the Pardee RAND Graduate School. The faculty committee that supervised and approved the dissertation consisted of Jim Hosek (Chair), Paco Martorell, and Jennifer McCombs.

RAND PARDEE RAND GRADUATE SCHOOL

Published 2014 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org/
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Email: order@rand.org

Table of Contents

Figures ... v
Tables ... vii
Abstract ... ix
Acknowledgements ... xi
1. Does Information Change Voter Behavior? An Analysis of California's School Board Elections ... 1
Abstract ... 1
1.1 Background on California School Board Elections and Accountability ... 6
1.1.1. California School Board Elections ... 6
1.1.2. Standardized Tests and Accountability in California ... 9
1.2. Information and Election Timing Framework ... 10
1.2.1 Retrospective Voting and Accountability/Information ... 11
1.2.2. Election Timing ... 12
1.3. Data ... 14
1.4. Basic Empirical Model ... 17
1.5. Results ... 18
1.5.1. The effect of PSAA on incumbent reelection rates in odd years ... 18
1.5.2. The effect of PSAA on incumbent reelection rates in even years ... 19
1.5.3. Robustness checks ... 21
1.5.4. Decomposition of the treatment effects ... 23
1.5.5. The effect of print media on the implementation of PSAA ... 24
1.5.6. The effect of newly elected school board members on API ... 26
1.6. Conclusions ... 28
2. Can recruitment and retention bonuses attract teachers to low performing schools? Evidence from a policy intervention in North Carolina ... 46
Abstract ... 46
2.1. The structure of the Mission Possible program ... 50
2.1.1. Recruitment, Retention, and Performance Bonuses ... 50
2.1.2. Staff Development ... 52
2.1.3. Performance Accountability ... 52
2.1.4. Structural Support ... 52
2.2. Data and methods ... 53
2.3. Results ... 59
2.3.1. Description of observable high school teacher changes ... 59
2.3.2. Difference-in-difference model of Mission Possible ... 60
2.3.3. Difference-in-difference specification check ... 62

2.3.4. The entry and exit of Algebra I teachers 63
2.4. Conclusion and policy implications 64
References 67
3. Can principal and teacher bonuses improve student achievement? Evidence from a policy intervention in North Carolina 81
Abstract 81
3.1 The structure of the Mission Possible program 85
3.1.1 Recruitment, Retention, and Performance Bonuses 86
3.1.2 Staff Development 87
3.1.3 Performance Accountability 88
3.1.4. Structural Support 88
3.1.5 Comparison of Incentive Design with Prior Random Control Trial Research 88
3.2. Data and methods 90
3.3. Results 94
3.3.1 Simple difference-in-difference model of Mission Possible on Algebra I test scores 94
3.3.2. Year-by-year difference-in-difference specification 95
3.3.3. Variation in the program's impacts 96
3.3.4 Evidence that the size of the recruitment/retention bonus matters 97
3.4. Discussion and policy implications 98
References 101

Figures

Figure 1.1: Turnout rate among registered voters from 1995-2006 35
Figure 1.2. Incumbent reelection rates by odd election years since accountability reform 36
Figure 1.3. Incumbent reelection rates by even election years since accountability reform 38
Figure 1.4 – Articles on the “Academic Performance Index” in California 43
Figure 2.1: Average percent of teachers with three years or less experience over time by whether the school is a Mission Possible school. 73
Figure 2.2: Average percent of teachers fully licensed over time by whether the school is a Mission Possible school. 74
Figure 2.3: Average turnover rates over time by whether the school is a Mission Possible school. 75
Figure 2.4: Distribution of teacher value-added for MP and non-MP schools, before and after the implementation of MP. 76
Figure 2.5 Difference-in-difference estimates of MP on standardized teacher effectiveness by year with clustered standard errors and placebo confidence intervals................................. 78
Figure 3.1: The estimated impact of Mission Possible reform on Algebra I standardized test scores... 109
Figure 3.2: The estimated impact of Mission Possible reform on black students’ Algebra I standardized test scores.. 110
Figure 3.3: The estimated impact of Mission Possible reform on free/reduced price students’ Algebra I standardized test scores... 111
Figure 3.4: The estimated impact of Mission Possible reform on English I standardized test scores... 112

Tables

Table 1.2: Difference Estimates of the Effect of Public Schools Accountability Act on Incumbent Reelection Rates During Odd Election Years 37
Table 1.3: Difference Estimates of the Effect of Public Schools Accountability Act on Incumbent Reelection Rates in Even Years 39
Table 1.4: Robustness Checks of Estimates of the Effect of Public School Accountability Act on Incumbent Reelection Rates in the even years 40
Table 1.5: Difference Estimates of the Effect of Public Schools Accountability Act (PSAA) on Incumbent Reelection Rates During the First Two Years of PSAA and for Uncontested Elections 41
Table 1.6: Difference Estimates of the Effect of Public Schools Accountability Act on Incumbent Reelection Rates 42
Table 1.7: Difference Estimates of the Effect of Public Schools Accountability Act (PSAA) on Incumbent Reelection Rates Year-by-Year 44
Table 1.8. Change in Academic Performance Index after the election of a challenger 45
Table 2.1: Incentive Structure for the Original Mission Possible Schools 71
Table 2.2: Student characteristics in analytic sample by school type 72
Table 2.3 Average and standard deviation improvements in teacher value-added at Mission Possible schools 77
Table 2.4 Difference-in-difference estimate at various teacher effectiveness percentiles 79
Table 2.5 The value-added of teachers entering and exiting MP schools, compared to non-MP schools 80
Table 3.1: Incentive Structure for the Original Mission Possible Schools 105
Table 3.2 Summary Statistics for Mission Possible (MP) and Comparison Schools in 2006 (the year prior to implementation) 106
Table 3.3 Summary Statistics for Mission Possible (MP) and Comparison Schools in 2006 (the year prior to implementation) using propensity score weights 107
Table 3.4 Estimate of the effect on Algebra I standardized scores from Mission Possible students 108

Abstract

It has long been thought that the United States education system is the great equalizer, lifting less advantaged children out of poverty and improving their chances for success in adulthood. The opportunity for economic and social mobility depends heavily, however, on access to high quality education. Recent research has raised concerns about degradation in the quality of schools serving higher-poverty neighborhoods: The achievement gap between low- and high-poverty students appears to have widened over the last quarter century (Reardon, 2011). In response to these concerns, federal, state, and local officials have enacted countless education reforms to improve the outcomes of low-income students. This dissertation examines two of those reforms to better understand how and if they are working.

The first paper focuses on California's state education accountability reform, which allowed the state to identify low-performing schools and target improvement efforts. The paper concentrates on a previously unstudied potential consequence of the reform: Whether the information on school academic performance, which had been previously unavailable, enabled voters to hold local leadership accountable. The results indicate that voting behavior depends on how well informed the voter is on school-related issues. Those who are less familiar with current school operations are more likely to utilize summary performance information—such as aggregate school test scores—to inform their choice of locally elected leader. Conversely, constituents who are more familiar with current school policies often discard student performance information. Taken together, these findings indicate that providing performance information to the public does not guarantee greater accountability of local education leadership.

The second and third papers assess a comprehensive reform to improve teacher and principal talent in high-poverty, low-performing schools. While the reform has various components, its main features are recruitment, retention, and performance bonuses for teachers and principals in schools with a greater concentration of high-poverty students. The second paper explores the intermediate outcomes of the reform by describing how it affected the movement of teachers within the district. The paper finds that the reform significantly improved the effectiveness of the teachers in the treatment schools. However, the improvements did not stem from attracting higher-performing teachers or removing low-performing teachers. Instead, the improvements were likely associated with a combination of effective leadership, improved working conditions, and performance bonuses. The third paper expands on these findings by exploring whether improving the talent within a school has an effect on student outcomes. Results suggest that while the program was not effective during the first three years of implementation, in the fourth year it shifted student performance by 0.28 standard deviations. These results provide evidence that targeted performance bonuses for teachers and principals could be used to improve academic performance in schools with large proportions of low-income students.

Acknowledgements

First and foremost, I would like to extend a huge debt of gratitude to my outstanding and dedicated dissertation committee: Jim Hosek, Paco Martorell, and Jennifer McCombs. While at the Pardee RAND Graduate School (PRGS), I experienced some truly challenging times, and through it all, they stuck with me. More importantly, they provided invaluable insight, comments, and edits to all three of these papers. The final product would not be what it is today without their assistance. I would also like to thank my outside reader, Katherine Strunk from USC, for her thoughtful and comprehensive comments on my papers.

On-the-job training was an important part of my PRGS learning experience; I took away much from these projects that I applied to my dissertation and will undoubtedly be able to apply to future work. Therefore, I would like to thank all those with whom I have worked closely over the years including: Jennifer McCombs, who gave me my first job; Paco Martorell, who is my mentor and whom I aspire to be like one day; Seth Seabury, whom I am glad to call a friend as well as a mentor; and Louay Constant, John Engberg, Robert Bozick, Gaby Gonzalez, Lindsay Daugherty, Gema Zamarro and J.R. Lockwood. I would also like to give a special thanks to Paul Heaton who taught two of the best classes I took at PRGS.

In addition, I wanted to thank my classmates, who were always there to discuss work- and non-work-related matters with me and who made my years at PRGS enjoyable. Adam Gailey, Chris Sharon, Helen Wu, Jack Clift, Sarah Kups, Vicki Shier, Debbie Lai, Amber Jaycocks, Aviva Litovitz, Susan Burkhauser, Matt Hoover, and Christina Huang were among those who helped me. Special thanks go to Sarah Kups, Adam Gailey, and Chris Sharon for reviewing and commenting on earlier versions of this dissertation.

Finally, I would like to thank my immediate and extended family for helping me through the ups and downs of the dissertation process. Judy and Maureen Manning provided invaluable child care so I could work on my dissertation while having an infant at home. Both my father and mother, Paul and Sherry Scherer, provided unwavering support that continues to propel me forward. Finally, and most importantly, is my loving wife, Jill Scherer. She is not only the reason that this dissertation is readable, but more importantly, encouraged me when things got tough and always celebrated my successes, no matter how small. Without her, I wouldn't be publishing this dissertation.

I gratefully acknowledge support from the Susan Way-Smith dissertation award. I would also like to thank the North Carolina Education Research Data Center (NCERDC) at Duke University for their assistance in providing the data on which I based two of my papers. I am fully responsible for the opinions expressed herein as well as any errors or omissions.

1. Does Information Change Voter Behavior? An Analysis of California's School Board Elections

Abstract

Government is increasingly allocating resources to measure the relative performance of providers of public goods, but the effect of this information on voting behavior is relatively unknown. In this paper, I study how voters react to performance information shocks in school board elections. I employ a difference-in-difference approach that compares incumbent re-election rates before and after the implementation of California's Public Schools Accountability Act (PSAA) in districts discovered ex post to be high and low performing, relative to their apparent peers. Using detailed district election data between 1995 and 2006, I find support that in low-turnout years, voters tend to disregard new academic performance indicators in favor of other forms of information. However, in high turnout years, voters use the academic performance metrics to hold incumbents accountable, though I find little evidence that school board members can affect the performance metric. Finally, I show suggestive evidence that despite relatively uniform election coverage in the media, as time elapses and more information becomes available, the incumbent re-election effects strengthen. Taken together, my findings contribute a more nuanced understanding to the expressive and retrospective voting literature.

Over the last decade, local, state, and federal governments have invested significant public dollars to create standardized metrics of comparison to aid policy-makers in their resource allocation and help constituents understand if providers of important public goods are working effectively and efficiently. Particularly since the passage of the federal law, No Child Left

Behind, the education sector has been particularly prolific in its use of standardized tests to compare academic performance across geographies. The availability of information about performance creates an opportunity for greater accountability of elected officials. For example, if a school district performs poorly on standardized tests relative to other districts in the state, presumably, voters can vote elected officials out of office. However, it remains relatively unknown empirically if voters or candidates use this public information and how its uses vary.[1]

A growing domestic empirical literature analyzes voters' responsiveness to information on political candidates.[2] In particular, several strong empirical studies examine how increased information improves voters' knowledge of a candidate and election-day turnout, and can have consequences at the ballot (see Healy and Malhorta, 2013 for a recent analysis). A parallel literature empirically investigates the importance of timing on electoral outcomes (Anzia, 2011; Berry and Gersen, 2011; Anzia, 2012). The authors of these studies find that switching from an off-cycle election[3] year to an on-cycle election year decreases elected officials' responses to dominant interest groups. Despite these advancements, it remains an open question how a "shock" in performance information affects voters with different electoral timing.

In this paper, I present new evidence on how a consistent signal of performance information—test scores of a school district—affects school board electoral outcomes. In

[1] An exception to this is Berry and Howell (2007) who examine the effect of performance accountability in South Carolina in a cross-sectional setting.

[2] There is a large international literature using information shocks to identify its effect on candidates and voters. These studies show that when information on performance or service quality improves, it can enhance government or program responsiveness to constituents, reduce corruption in the short run (Besley and Burgess, 2002; Reinikka and Svensson, 2005; Olken, 2007; Bojrkman and Svensson, 2009), and promote electoral accountability (Ferraz and Finn, 2008; Ferraz and Finn, 2009; Banerjee et al., 2010; Chong el al., 2010). Yet, other research has shown that information can increase candidate corruption in the long run (Bobonis et al., 2010). Candidates selected for random audits in previous periods, no longer fearing that they could be selected, actually increased their corruption beyond the initial levels.

[3] "Off-cycle election" refers to an election held on a different day than other major elections, such as a presidential or gubernatorial election.

particular, it takes advantage of panel data to examine how constituents and candidates respond to district performance information and how their composite reactions vary if the election is off- or on-cycle. The analysis utilizes the state's requirement, established under California's Public Schools Accountability Act (PSAA), to post publicly test score information, to rank schools of similar demographic make-up, and to impose sanctions on poor performing schools after 1999. There were no uniform reporting requirements or sanctions across the state for the four years prior to PSAA.

My research design exploits the fact that I can identify districts that are similar on variety of observable characteristics and that are discovered ex post – after PSAA takes effect – to be high and low performing, relative to their apparent peers. I then use a difference-in-difference model to identify the causal effect of this school performance information on incumbent re-election rates. I am particularly interested in whether these effects differ in on-cycle (even years) and off-cycle (odd years) election years. Based upon my theoretical framework, I expect to find a difference in the treatment of information between on- and off-cycle elections. This allows me to employ a triple difference model for the on-cycle elections where the third difference is between even- and odd-year elections. Finally, I perform supplemental analyses to better understand how the short-term effects of the program, the rerunning behavior of candidates, media coverage, and the ability of a school board member to influence test scores can inform my findings.

I find that in off-cycle years, test score information has no effect on voters' decisions. In fact, while not statistically significant, the sign of the effect indicates that voters might be deliberately placing little weight on information on test scores. One interpretation of this result is that information is only useful if it improves or reduces the noisiness of the voters' prior assessment, since it is likely that voters during these years are more directly involved with schools (e.g., teachers, parents in the parent teacher association). An alternative view is that self-interested groups represent a higher proportion of the votes in off-cycle elections and vote for their narrow interests, disregarding the performance metric.

In years with higher voter turnout, low test scores have a negative effect on the reelection rates in poorly performing districts. The results are robust to measuring test scores in levels as well as gains (though the effects for gains are smaller in magnitude, providing some evidence that voters care more about the absolute rather than improvements in performance). These results have a corollary conclusion to the odd-year results: Either voters do not pay attention to the day-to-day operations of schools and hence use summative metrics like the Academic Performance Index, or that the electorate is collectively less self-interested in this measure of school performance than those in off-cycle years.

Next, I try to further undercover the make-up of the effects I observe during the on-cycle years. First, I try to separate whether the effect I observe arises from the accountability reforms or mainly from the information requirement of the law by exploring the short-term effects of PSAA. As the accountability reforms took several years to take effect, the early analysis should highlight information's effect on voters. I find more than 2/3 of the overall effect's magnitude stems for informing the public. Yet, it seems more likely that the difference I observe between even- and odd-cycles is due to longer-term effects. Second, since the effect I observe is a composite of candidates' rerun and voters voting behavior, I examine the rerun behavior of incumbents. I find surprisingly little evidence that candidates incorporate the increase in information in their rerun decisions and the majority of the effect is driven by voters removing incumbents.

Finally, I explore two possible reasons for the effects that I observe. First, I perform a simple analysis as to assess whether school board members have much control over academic outcomes by examining the test score growth rates after the election of a new school board member. I find that when new members are elected to the board, school districts do not experience changes in performance. This result possibly indicates that a rational voter would throw out metrics like performance due to beliefs that a single school board member has little influence on test scores – such as I find during the off-cycle election years. Second, I present some correlational evidence that the more accessible and longer information is available on school, district, and state websites, the stronger the effects are between even and odd cycle years. This result holds despite relatively consistent levels of media coverage over the eight years post-reform.

However, unfortunately, both of these investigations for the underlying reasons suffer from data limitation. Ideally in additional to the detailed voting data, I would have reports of who teachers unions endorsed. This would enable to understand whether union strength is greater during the off-cycle elections, and corroborate what other authors have found and place my finding about new school board members in a more complete context. Further, it would be ideal to have information about website usage or even detailed local newspaper and non-print media coverage during on- and off-cycle elections rather than the crude tool of newspaper mentions from a large database. These limitations mean than future research must work to uncover the "black box" of why I observe these effects.

Overall, my results reflect a nuanced perspective regarding the role of information on political accountability. In particular, more information does not always mean that it will affect voter's choice. Yet, since much of the prior literature's elections are coordinated with national and state-wide elections, my results are consistent with prior research that information enhances political accountability during major election cycles. Voters tend to punish incumbents in school districts that perform poorly on the state accountability scale. At the same time, I also raise some doubts whether standardized test scores are the correct metric that should be used to evaluate the performance of school board members. This question should make policy-makers cautious when designing information accountability system. Finally, my paper lends some support to the value of websites and graphics that allow districts, cities, and the federal government to illustrate their performance for those who tend to incorporate the information. These simple visuals (assuming they gain viewership) could be an important check, beyond the media, on electoral power.

This paper proceeds in six sections. The first section provides background on school board elections in California and the state's Public Schools Accountability Act. The second section provides a framework to assess my results. The third section describes the data. The fourth section lays out my empirical model. Fifth, I review the results, and finally, I conclude.

1.1 Background on California School Board Elections and Accountability

1.1.1. California School Board Elections

School board elections provide a unique environment to test theories of information and accountability. While federal and state education departments mandate some policy decisions, most of the power to determine the educational vision, goals, and policy for a district resides at the local level. Much like corporate boards, the school board focuses on the ends, while their elected official, the superintendent (the CEO of a school district) focuses on the means. The board's objective is to align the goals, standards, and strategy with the budget of the school district. They are also responsible for using data and public testimony to assess their progress towards these goals and make corrections as are needed. One of their key metrics to evaluate school performance in the modern day, are standardized test scores. Finally, unlike other locally elected officials, like a mayor, school boards central focus is on a single public good: a student's education. This combination of a standardized metric of performance and the primary function of providing each student with the highest quality education means school board elections are an excellent stage to assess my research questions.

California is an interesting landscape to assess school board elections because of its diversity. School districts in the state can contain only elementary schools (K-8), only high schools (9-12), or they may be "unified" and include elementary and high schools. In the 2011-12 school year, there were 540 elementary, 80 high school, and 338 unified school districts.[4] Even among these broad classifications, school districts can range significantly in population size and diversity. About 10 percent of all districts are "very small" (serving fewer than 100 students) according to a Legislative Analyst Office report in 2011.[5] Moreover, 230 of the state's districts contain only a single school. These "small districts" can be contrasted with Los Angeles Unified School District (LAUSD) that manages 937 schools and serves over 650,000 students.[6] Furthermore, while LAUSD's student population is 90 percent non-white, Laguna Beach Unified's non-white population is 18 percent.[7] This heterogeneity enhances the generalizability of my findings to other states.

The timing of California's school board elections is fairly unique, and recent evidence has shown that timing of elections is important. California is one of eight states with a mixture of

[4] California Department of Education, http://www.ed-data.k12.ca.us/Pages/Home.aspx

[5] California Legislative Analyst Office, http://www.lao.ca.gov/reports/2011/edu/district_consolidation/district_consolidation_050211.aspx

[6] California Department of Education, http://www.ed-data.k12.ca.us/Pages/Home.aspx

[7] Ibid.

on-cycle and off-cycle elections (Anzia, 2011). While a considerable political science literature examines the effect of non-presidential and off-cycle elections on voter turnout,[8] a much smaller and emerging literature has examined some of the policy and voting consequences of these off-cycle elections (Dunne, Reed, and Wilbanks, 1997; Trounstine, 2010; Anzia, 2011; Berry and Gersen, 2011; Anzia, 2012). Many of these studies focus on teachers unions, the largest and most powerful interest group during these elections (Moe 2006; 2009). They find that changing the timing of voting translates into much lower voter turnout, which favors organized interest groups (Anzia, 2011; Berry and Gersen 2011; Anzia, 2012). The lower turnout allows teachers' unions to wield greater power on the elected board, translating into higher teacher compensation.

For my analysis, I utilize the difference in voter turnout between elections during even and odd years. In California, almost 40 percent of school board elections in my data occur in "off-cycle" years (e.g., 1999). Figure 1 shows the turnout rate, defined by the total number of school board votes in my sample divided by the number of registered voters reported on the California Secretary of State website.[9]

Similar to the findings of the previously mentioned literature, I find that voter turnout is lower during off-cycle elections. Turnout rates are about a third to a quarter lower in the off-cycle years when compared with the even years. With one exception, California's special election for governor in 2005, voter turnout in odd years is consistently lower than 20 percent of registered voters. As others have found, this pattern leads me to believe that the nature of these elections are different. During the odd election years, it seems likely that voters are more likely to be parents actively engaged in their children's schools and/or teachers and administrators

[8] Although not complete, see Wolfinger and Rosenstone, 1981; Patternson and Caldeira, 1983; Hess, 2002.

[9] Note that these participation rates differ from Berry and Gersen (2011) because they used voting-age residents rather than registered voters as reported by the state. In addition, they need to throw out several important large districts from their data, such as Los Angeles Unified School District and San Francisco Unified School District.

directly involved with schools. For these odd-year voters, a comprehensive score like the Academic Performance Index may not be as informative since they may be closely involved in the day-to-day activities of the school/district, or as some have argued, may care more about an interest group's endorsement than performance of the district (Moe 2006; 2009; Anzia, 2011; Anzia, 2012).

1.1.2. Standardized Tests and Accountability in California

Although California has a history of measuring students' academic performance, the state lacked a common measurement system for several years leading up to accountability reform. From 1990 to 1995, California used the California Learning Assessment System (CLAS) to measure students' academic outcomes. However, after poor performance on the National Assessment of Educational Progress, sometimes referred to as the nation's report card as it provides rankings across states, CLAS was discontinued. School districts then selected their own tests, making it impossible to compare student performance across the state. In an effort to remedy this problem and enable reward of "good" schools and punishment of "bad" schools, California enacted the Public Schools Accountability Act (PSAA) in April 1999. PSAA has three main components: the Academic Performance Index (API), the Immediate Intervention/Underperforming Schools Program (II/USP), and the Governor's Performance Award (GPA).[10]

By identifying how schools were performing, the state could reward good schools, and provide additional monetary assistance to schools that consistently fail. Schools that continually miss targets become labeled as II/USP, which requires them to offer additional tutoring for

[10] 2008-09 Academic Performance Index Report: Information Guide. http://www.cde.ca.gov/ta/ac/ap/documents/infoguide08.pdf

students, provide options for sending students to an alternative school, or close programs down entirely. The law also created a "similar school ranking" by identifying 100 school cohorts of similar schools based upon observable characteristics and ranking them by deciles. Given the circumstances in 1998, the introductory API scores represented a significant increase in the information available to voters on the performance of school districts.

For this study, I focus on the information conveyed in the API score. The API is a numeric index between 200 and 1,000 and reflects a school's performance across all standardized tests administered by the state. As a composite index, it combines into a single score the results of all the tests from the Standardized Testing and Reporting Program (e.g., math, English and subject tests) with the California High School Exit Exam. Depending on where a school falls on the scale, a certain amount of growth is required every year to comply with PSAA. For example, if a school is between 200 and 690, the required growth to pass is five percent of the difference between the school's score and 800. The first baseline API scores were collected in 1999, and schools were subject to accountability in 2000.

1.2. Information and Election Timing Framework

The following framework is to clarify some of the underlying incentives and policy results for an information shock when the timing of an election varies. In particular, I briefly discuss the retrospective voting literature. Then I use a simple cost and benefit model of voting behavior to show how the median voter changes based upon the specific election. The cost-benefit framework is based upon the work of Dunne, Reed, and Wilbanks 1997 as well as the theoretical model developed by Banerjee et al., 2010.

1.2.1 Retrospective Voting and Accountability/Information

Retrospective voting, how citizens respond to their perception of government performance, plays a key role in democratic accountability (Besley, 2006; Ashworth, 2012). Until recently, much of the empirical work has focused on economic retrospective voting – the use of economic performance indicators to hold candidates accountable (Anderson, 2007). However, there are several reasons to be concerned about the validity of the observed effects of economic retrospective voting including, but not limited to: 1) the extent to which a politician's fiscal policy can influence the larger economy, 2) what are the "right" economic indicators to utilize, and 3) the right counterfactual – what the economy would have been like under a different candidate (Healy and Malhotra, 2013). To address some of these concerns a recent portion of the literature has begun to explore non-economic indicators, such as test scores, disaster response, and war causalities (Berry and Howell, 2007; Healy and Malhotra, 2009; Bechtel and Hainmueller, 2011; Karol and Miguel, 2007; Kriner and Shen, 2007; Grose and Oppenheimer, 2007). These studies generally find that voters respond to positive and adverse information about candidates.

This paper builds upon this literature by using test score performance to understand retrospective voting in school board elections. As noted above, a school board's central goal is to produce a high quality education for students and it can be evaluated using a standardized metric: test scores. After the introduction of accountability in California, my data allows me to identify districts that appear similar based upon observables characteristics, but differ based upon their students' performance. If voters use this information to hold candidates accountable, we would expect to see differential incumbent reelection rates between "low" and "high" performance districts.

It's important to note here that up until this point, that I have used information and accountability interchangeably, since an important component of the accountability reforms was providing information to the public. However, accountability, sanctions and rewards for performance of a district/school, itself could be driving any effects that I observe. However, since the reasons and actions items for low performing school improvement were not due to the state until January 2002, it is possible to separate the effect of accountability and information empirically during the early years of the program.

1.2.2. Election Timing

To understand how the median voter might change for different types of elections (e.g., on- or off-cycle), first consider a single jurisdiction comprised of different voter types utilized in Dunne, Reed, and Wilbanks (1997). To illustrate the different interest groups, the authors consider a cash transfer allocated by school boards. The first group is teachers and administrators who receive a direct benefit from the resource transfer, presumably in the form of salary and benefits increases. The teachers and administrators are the smallest fraction of registered voters. The second group is parents who experience an indirect benefit from the resource transfer through enhanced quality of education for their children.[11] Parents make-up a larger portion of the electorate than teachers. The final group is the general electorate that experiences positive externalities from high-quality education through reduced crime and other outcomes. These voters are largest group among registered voters.

I now consider these types of voters within a cost-benefit framework. Consider a benefit $B_i(n, e)$, which is a function of n, the number of voters and e, the number of elections on the ballot. The benefit is indexed by i, the type of voter. The benefit is decreasing at a decreasing

[11] Note that empirically, the relationship between school expenditures and education quality has been shown to be weak (Hanushek, 1997). However, for expositional purposes, it makes sense to highlight the connections.

rate in n and increasing in e. Benefits can be ranked by the type of voter with teachers and administrators receiving the largest benefit and the general electorate receiving the least based upon the example in the prior paragraph. The changes in type can be thought of as an intercept shift in the benefit. Based upon this ranking, different groups have different incentives to turnout, holding the number of voters constant. There is a fixed cost, $\bar{v}$, of voting in an election.

Next I consider how these costs and benefits differ from year-to-year. The cost of attending the polls is fixed because staying informed, etc., does not vary between an even and odd years. In other types of elections it might be more costly to acquire information during the off-cycle elections, however based upon a media scan in Lexis-Nexis of the term "school board," and "California" I find little difference in the print media coverage of school boards. Furthermore, there is virtually no television and radio transcript coverage of school board elections in Lexis-Nexis indicating that this cost assumption is reasonable. In contrast, the benefit of voting does vary from year to year. As there are more elections on the ballot during even years (e.g., presidential as well as school board elections) the benefits are higher for everyone during these years. Using the formulation above we can write a simple turnout equations where it must be true that $B_i(n, e) > \bar{v}$. Considering this equation, those with the highest benefits (e.g., teachers and administrators) will turnout in years when the number of elections on the ballot is small (e.g., only the school board is on the ballot). Presumably this logic helps to create the pattern noted in Figure 1. Based upon this model, the median voter will be closer to the administrator and teacher group when the timing of the election includes fewer elections on the ballot.

Continuing with the example of the three voter types, we can also understand how useful standardized test score would be. Teachers and administrators are likely the most informed voters, while the 18 year old with no children in the school system is unlikely to follow the day-to-day operations of the school. Thus, a sudden increase in the amount information about a school will be least useful to those with the most information, teachers and administrators. An important parallel argument is special interest groups, teacher unions, might be the most likely to vote their interest over objective information on schools (Moe 2006; 2009; Anzia, 2011; Anzia, 2012). Both of these will result in the same observed outcome in my data, that voters with a median closer to teachers and administrators will be more likely to disregard objective information shocks in elections. Conversely, those voters who are least informed are more likely to utilize publically available information for their election decision.

Within the context of my data, I use this framework to make several predictions. Based upon the framework developed above, when there is the smallest benefit (odd-year elections with only the school board on the ballot) only those most motivated turn out. These voters will be more likely to be part of interest groups (e.g. teachers, parents) who monitor decisions of the school boards more closely. Thus, empirically, we would expect to see the odd-year voters disregard the API. During even years, Figure 1 shows electoral participation is relatively high and likely shifts the median voter towards the general electorate because the benefits of having a national- or state-level office on the ballot increases the voting benefit. However, this means that the median voter knows less about their local school board elections. If this is the case, empirically we should observe that voters are more responsive to the API.

1.3. Data

The data used for this analysis is a combination of three large datasets. First and most importantly, I use the California Election Data Archive (CEDA) compiled by California State University, Sacramento to analyze voting behavior. CEDA has detailed local election results, including election date and office, as well as the number of votes received by each candidate running for office for all state elections and ballot initiatives in California from 1995 through 2006.[12] From this dataset, I kept all school board elections held during the period.

I merged these data with detailed information from the California Department of Education (CDE) API data files. These data files not only include the API scores for 1999-2006, but other school characteristics, including the California School Character Index (SCI) and its components. The SCI calculates predicted API scores by taking into account important characteristics of schools that might make them higher or lower performers.[13] The SCI uses characteristics, such as pupil mobility[14], ethnicity, socioeconomic status, percent of teachers who are emergency and fully credentialed[15], percent of English language learners, average class size per grade level, and type of school year (e.g., all year round or nine months), to estimate a single index. All the school-level data, including the SCI and API are aggregated up to the district level by weighting each school by its proportion of district enrollment.

Finally, in order to include district demographics prior to 1999, I use the National Center for Education Statistics Common Core of Data (CCD). The CCD collects basic racial composition

[12] There were some inconsistencies with the data. Sometimes the sum of the individual votes did not equal the total, the percentage of the vote share did not match the percentage reported, or one of these three variables was missing. In these cases, I imputed the variable based upon the other available variables. In all cases, at least two of the three variables needed to calculate the third were available.

[13] The SCI uses a regression-based method to construct the index, by regressing the API score on factors including: pupil mobility, pupil ethnicity, the education level of parents, the percentage of free or reduced price lunch in the school, percentage of students classified as English language learners, if the school is multi-track or a year-round school, average class size, and percentage of credentialed teachers. After estimating the regression, the actual school characteristics are multiplied by the regression coefficients to yield the index value. There are a handful of other small augmentations. For further detail on the process, see the document available on the California Department of Education website: http://www.cde.ca.gov/ta/ac/ap/documents/tdgreport0400.pdf. The SCI is then used to create cohorts of similar schools and the state reports a similar school rank. The similar school rank is not used for accountability purposes.

[14] “Pupil mobility” is defined as the percent of students who first attended the school.

[15] In the teaching profession, in order to teach, it is required that the teacher be certified by the state. However, in cases where there are no certified teachers available, districts do allow “emergency credentialing” that allows someone without a credential to teach in the classroom.

and the percentage of free/reduced price lunch students (a proxy for poverty) for all schools in the United States.

To construct treatment and comparison groups, I consider a "slightly sophisticated voter" who examines the API scores of similar school districts in the state to compare against his district's API score. I elect this method because being able to rank similar schools is an important part of PSAA.[16] To accomplish this, I use the SCI index. I rank ordered the SCI and matched each district to its closest in composition, as measured by SCI, without replacement. Within each pair, I noted which school had the worse API score and designated them in the treatment group. Thus, the post-period voting record in the revealed higher performing school districts provide the estimate of the "counterfactual"—similar districts with better performing school board members.

Because parents needed to make decisions about locating in school districts based solely on observable factors, it seems reasonable that prior to the test score release in 1999/2000 they would view two districts on the SCI index as similar in quality.[17] Then the 1999/2000 API score provides a strong signal of quality.

The top panel of Table 1 reports observable characteristics of high- and low-performing districts for even and odd years using my preferred method to assign the groups: the SCI. Based upon the p-values reported in columns (3) and (6), the observable characteristics of the two samples, with the exception of the API score, are not statistically different from one another. Of course, by design, the 1999 API scores differ significantly, where the high-performing group consists of higher performing districts.

[16] Note that I consider other methods in my specification checks in the results section.

[17] The exact weighting by parent of certain district/school characteristics could be different than the weights in the calculated SCI (e.g., only placing weight on school race or percent free/reduced lunch). However, absent knowing these weights, the SCI is a good approximation.

In the lower panel, we see that prior to the state accountability standards, there is no difference in the incumbent rerun and re-election rates except in the odd years the high performing group is a little more likely to rerun. This could potentially bias the outcome downward, which I will address in my results section.

1.4. Basic Empirical Model

To analyze whether I observe retrospective voting for the California school board elections I use a difference-in-difference estimation strategy. In particular, I am interested if after the introduction of PSAA whether the incumbent reelection rates change differently for the relatively lower performing district. To estimate this I use the following equation:

$$v_{it} = \beta_1 lowperform_{it} * post_t + \beta_2 lowperform_{it} + \beta_3 post_t + \beta_4 X_{it} + t_t + d_i + \varepsilon_{it} \qquad (1)$$

Where v_{it} is incumbent re-election rates for district i in period t, $lowperform_{\text{it}}$ is an indicator for a lower performing district, $post_{\text{t}}$ is an indicator which is 1 for all years after 1998, X_{it} is a vector of observed district covariates[18], t_{t} is a vector of year fixed effects, and d_{i} is a vector of district fixed effects. It is important to note here than any unobserved factors not accounted for by matching on SCI will be accounted for with the district fixed effects. The key coefficients of interest will be β_1, which will indicate the percentage point change in the incumbent re-election variable due to the implementation of PSAA identify which districts with similar observable characteristics are higher or lower performing. I calculate Bertrand et al., (2004) standard errors by block bootstrapping (i.e., clustered at the district level) 400 samples. For further reference on this technique, see Bertrand et al. (2004).

[18] Note that this set of covariates is less complete than those in the API files since I use the Common Core data for these covariates. It consists of measures of poverty and race within the district.

1.5. Results

1.5.1. The effect of PSAA on incumbent reelection rates in odd years

Figure 2 plots reelection rates on the vertical axis and the horizontal axis displays the calendar years before and after the implementation of PSAA (with the year of PSAA denoted year 0) for the odd years (e.g., year 0 equals 1999). The vertical line in this graph separates the pre- and post-implementation periods. Although these numbers are not regression adjusted, the figure demonstrates a graphical form of the basic difference-in-difference model. This simple graph does not allow for a thorough analysis of PSAA in the odd years, but it does show preliminary evidence that the high- and low-performing groups tend to mimic each other fairly closely indicating there is no effect of the revealed performance information for these voters.

Table 2 presents several specifications for the difference-in-difference models during the odd election years. In column (1) I present the simple difference-in-difference model shown graphically with year fixed effects in the model. We cannot distinguish the coefficient from zero. While that is the case, the coefficient sign takes a positive value indicating that lower performing incumbents are *more* likely to be reelected in periods after the release of additional information. Column (2) adds key controls and column (3) adds district fixed effects to the model, but this does not change the significance and only increases the magnitude of the coefficient. Graphically we can see that this effect might be largely driven by the 2005 election (three odd election cycles after accountability reform), which as noted above was unusual because of the special election held that year. Column (4) reruns the specification in column (3) excluding 2005 from the analysis. While this does reduce the magnitude of the coefficient, it remains positive and not statistically different than zero. Thus, the fact that incumbents in worse performing districts tend to have higher reelection rates after the implementation of PSAA seems to be robust to various specifications.

In line with the theoretical prediction, these results provide preliminary support that odd-year voters are throwing out the information on API score although the effect is not statistically different than zero. The fact that we can't distinguish the effect from zero means the odd years provide an important comparison group to the even years that I will discuss further in the next section.

1.5.2. The effect of PSAA on incumbent reelection rates in even years

The way to test if there is a statistical difference between even and odd years, would be to compare (1) high- and low-performing districts, (2) before and after the implementation of PSAA, and (3) between even years and odd years. The best model to test this would be a difference-in-difference-in-difference strategy. Now our new empirical model equation is:

$$\begin{aligned} v_{ite} = {} & \beta_1 lp_{it} * post_t * even_e + \beta_2 even_i * post_t + \beta_3 even_e * lp_{it} + \beta_4 lp_{it} * post_t + \\ & \beta_5 even_i + \beta_6 lp_{it} + \beta_7 post_t + \beta_8 X_{it} + t_t + d_i + \varepsilon_{ite} \end{aligned} \tag{2}$$

where e indexes even or odd year and even_e is an indicator variable if it is an even or odd year. I have abbreviated the lower performing variable to lp_it. The other variables in equation (2) are similar to those discussed in equation (1). In this case, the three-way interaction term between lower performing, post and even represents the estimate of PSAA effect in even years on the incumbent's reelection rates.

Prior to assessing the third difference it is important to see graphically if there is a relationship between the high- and low-performing group in the pre- and post-intervention time period. Figure 3 shows the relationship between incumbent reelection rates and the number of election years since PSAA for high and low performing districts. This figure is parallel to Figure 2.

We want to be caution once again, since these are unadjusted numbers, however, prior to PSAA we see that the treatment and comparison group have relatively parallel trajectories, the key identifying assumption for a difference-in-difference model. However, after the implementation the treatment group experiences a rapid decline in reelection rates. The lower rate persists throughout the three even post-intervention election cycles.

The first column of Table 3 presents the even year difference-in-difference model using specification (3) in Table 2. As expected, based upon Figure 3, we see a significant negative effect of PSAA on being reelected to office in the post-period. These results confirm that voters during the even years are voting lower performing districts out of office at a higher rate than the their higher performing counterparts in the post-PSAA period. Thus, voters are holding incumbents accountable during the even years.

The second column shows the triple difference estimate with controls, years and district fixed effects in the model. This increases the magnitude of the effect and is also statistically significant at the 5 percent level. The increase in magnitude clearly shows that when compared to the odd-year electorate, the even-year voters are much more likely to use API scores to hold incumbents accountable. These results show that PSAA had an effect on incumbent reelection rates compared to the off-cycle years. Since the mean reelection rate is 65 percent, a 13 percentage point reduction is almost a 20 percent decline in the reelection rates of incumbents. Thus, the release of information had a practical effect on the incumbent election rates.

Up till this point of the paper, we have been examining the API levels as the key criterion voter use to assess school board performance. However, as mentioned in the data section, for most school districts with scores under 800, the PSAA law holds schools accountable based upon the growth of the scores from the prior year. In addition, the growth score is a better metric for the efficacy of the board since it demonstrates how much things improve year-over-year. The growth score is also reported by most districts along with the level score. The third column of Table 3 replicates the difference-in-difference specification in column (1) for the gains in even years. The difference-in-difference model provides evidence that while voters react to the information in the direction we expect based upon the pervious results, because we can't distinguish this effect from zero, it appears voters do not focus on gains as much as the overall API. However, when compared to the voters of the odd years, the magnitude of the estimate is similar to the levels model. Once again we see that the DD estimates show that in the post-PSAA period voters reelected incumbents at a differential rate in the lower performing districts. But when compared with the off-cycle electorate, it is clear that the two groups use the API differently from one another. The gains DD and DDD results provide some qualitative evidence that while the gains a district makes in API are important; the focus of voters is on the absolute score.

1.5.3. Robustness checks

Table 4 presents several alternative specifications to assess the sensitivity of the estimates. First I construct an alternative treatment groups. In specification 1, I assume a less nuanced voter who does not care about the demographics and other factors of the district and only cares if the district is high or low performing. To accomplish this specification, I assign the lowest quartile of school districts as my treatment group and the remaining sample is the comparison group.[19] These results show weaker, but similar effects as Table 3 for the levels model. I repeat this for the gains score; however, small gains could be because it is hard for a high performing district to improve. This fact means that the lower quartile of gains lumps districts across the

[19] One quarter is fairly arbitrary, thus I varied the break point by +/- 20 percentage points with qualitatively similar results to those presented here.

performance spectrum into the treatment group. With this understanding, it is reasonable that the gains specification should be interpreted differently than they are in Table 3.

Second, rather than identify a treatment and comparison group based upon SCI or API, I incorporate both metrics, API-SCI, into my model directly. The new variable is a continuous treatment measure that identifies the performance of a district (API) above or below its expected performance based upon the characteristics of the district (SCI). I then plug my new measure in place of β_1 in equation (1) and (2).[20] In this case rather than a negative effect, we hope to observe a positive relationship between the continuous measure and incumbent election rates. This effect would mean that as a district improves its performance over what is expected, incumbent reelection rates should increase. In fact, in specification 2 in Table 4, this is what we observe for the even-year difference-in-difference. The point estimate indicates that for a one point improvement in API above the SCI increases incumbent reelection rates by 0.4 percentage points though it is only marginally statistically significant. A one standard deviation in API-SCI is about 50 point which would increase reelection rates by 20 percentage point. For the triple difference, the observed effect is not statistically different than zero, though like the estimates in Table 3, it is double the size of the difference-in-difference specification.

A third concern to the observed effect is that many of the school board elections are uncontested (i.e., close to 40% of my sample). Uncontested elections build in a mechanical relationship in my data such that regardless of the API score, the incumbent will be elected. To address this potential concern, I include a dummy variable in the equation indicating whether or not the election was contested. The results are in the third row of Table 4 show that while

[20] For equation (2), I plug in the continuous measure and multiple it by even and include all the underlying interactions.

uncontested election slightly reduce the magnitude of my estimates, compared to Table 3, they have a minimal effect overall.

Finally, as noted in Figure 1 and 2, 2005 was an unusual election year and there is some indication that it skews the results. As such, I run a specification without 2005 or 2006 in the model to see if it changes the result. This specification does not substantively change the results in Table 3. Taken together, my results are robust to alternative specifications of the estimates.

1.5.4. Decomposition of the treatment effects

Until this point in the analysis, I have combined the effect associated with the additional information required by PSAA and the accountability mechanisms that it put in place. To disentangle these effects I employ an additional strategy. While the information on test scores were available immediately, the sanctions associated with PSAA were not fully put in place until several years after the implementation of PSAA. Thus, if I observe an effect, especially in the difference-in-difference model for the early years of PSAA, it will provide evidence that the effect is due to information rather than accountability. The first column of Table 5 shows the even-year difference-in-difference estimator in Table 3 column (1) where I exclude 2001-2006 from the model (i.e., the only post-PSAA year is 2000). While not statistically significant, the magnitude of the effect is more than 2/3 of the size of that observed in Table 3. As such, there is evidence that potentially a large portion of the observe effect in Table 3 is related to simply the provision of information. However, the second column of Table 5 shows the triple difference effect for the same time period. Here we observe the magnitude makes up a much smaller portion of the total effect in Table 3. Yet, the triple difference specification measures something slightly different than the difference-in-difference model. It measures how different types of voters use the information/accountability. It is less likely that I would observe an effect of accountability using the triple difference between even and odd-year because accountability should not affect schools differentially based upon the timing of their election. Thus, even though the difference in the earlier years of PSAA is relatively small compared to the overall effect that I observe in Table 3, I still believe the effect is mainly due to information rather than accountability.

Second, an important piece of the analysis is not only whether voters react, but do candidates react strategically to the information. The effects that I observe in Table 3 are a composite of candidates deciding not to rerun because of PSAA and the incumbents being defeated. To better understand where the effect stems from, I perform a separate analysis of the candidate rerun behavior using similar models. Table 7 replicates the even- and odd-year analysis for the difference-in-difference model as well as the triple difference. While the signs of these effects go in the same direction as what we observed in the Tables 2 and 3, the magnitudes are too small to differentiate them from zero. However, the magnitudes of the coefficients tend to be about 1/3 to 1/2 of the size of the total effect. Therefore, some of the effect can be explained based upon incumbents declining to rerun for office.

1.5.5. The effect of print media on the implementation of PSAA

Berry and Howell (2007) observe an initial effect of information in South Carolina's 2000 school board elections that fades away in the 2002 and 2004. They explore this pattern and conclude that a significant change in the type of news media coverage over time combined with less precise measure of school performance lead to this result. While the next results are based on a correlation, I observe a different pattern in California.

First I scan large and local newspapers in California for the two months prior to school board elections using the key search terms "school board" within the Lexis-Nexis database.[21] In the two months prior to the election this amounted to approximately 1100-2500 articles which given the size of the state and the number newspapers included in the search, is fairly minimal coverage. Within each of these searches I added the key term "Academic Performance Index." For the years of interest between 2000[22] and 2006, I observe that the API is mentioned in about 1 to 2 percent of the articles consistently across the years, with little difference between even- and odd-cycle elections--a uniform distribution rather than the declining one found by Berry and Howell. These results would imply that we would see consistent effect of the information across the years after the intervention. Figure 4 shows a second test where I examine more broadly in California newspapers the number of times the term "Academic Performance Index" appears. Here we see again, that while there is a large jump in 1999 and 2000, the coverage is relatively consistent in the subsequent years. A secondary point to the graph is that despite there being over 700 districts in my sample and over 1,000 in California; we see in Figure 4 that the API is mentioned less than 300 times in the two month prior to the election. Thus, among school coverage, mentions of API are relatively infrequent.

To correlate the newspaper coverage findings with my data, I employ a third empirical strategy. I continue to use a triple difference to measure the difference between the on- and off-cycle election, but instead of using the pre- and post-intervention indicator (e.g., $post_i$), I multiply the even*treatment interaction by individual year dummies, where the omitted category is one year prior to the intervention. Table 5 shows the three-way interaction by year from

[21] LexisNexis does not compile all local newspapers in California, but tracks many of the major newspapers in various regions of the state. This poses a limitation on the analysis.

[22] There was little coverage of the Academic Performance Index in 1999, as can be seen in Figure 4, thus I exclude it from the analysis

intervention. As this requires considerably more covariates in the model, I only find results statistically different than zero in the final year of my data. However, looking at the point estimates we see that in the post-intervention period, the magnitudes of the coefficients are increasing in size. These results conflict with the simple newspaper analysis in Figure 4. The juxtaposition of the raw media attention on the API and the steadily increasing difference between API usages between on- and off-cycle voters indicates that there is something else happening at the same time. It could be that information presentation improved during this time period and while the usage of API increased for on-cycle users, off-cycle users continued to use other sources of information. While I don't have direct evidence from saved web sites, the data websites for API certainly improved in their data availability as well as their sophistication over time. It seems reasonable that as districts and schools became better at communicating performance to constituents, the effect of the information strengthened over time. Alternatively, it could be the case that as teachers' unions became more opposed to standardized test scores, they increasingly rejected API as the performance metric over time. However, to investigate this research question, a more thorough analysis with superior data to mine must be performed.

1.5.6. The effect of newly elected school board members on API

One possible reason that the low turnout years uses information about performance differently than the high turnout years is that individuals could be skeptical of school board members' ability to change API. An easy test for this possibility is to examine how much a school improves after the election of a new school board member. To accomplish this, I limit my sample to 2000-2006, years where gains in the API could be measured. I created an indicator whether any new school board members were added during a particular cycle. Then controlling for the demographics in the SCI index and the lag of student test scores, I examine if there were any gains either two years or four years after the new school board members were elected. I chose two and four years after the election cycle, since some members of the board (not the individual) would be up for re-election two years afterward. After four years, the newly elected school board member would be up for re-election again.[23]

The results of the analysis are shown in Table 6. The first column examines if there is any effect on all elections held 2-years after the initial one. While the magnitude indicates there is actually, on average, a single point loss in the scaled score when a new school board member is elected, this can’t be differentiated from zero. To understand what is driving this relationship better, in columns 2 and 3 of Table 6, I break apart the effect into even and odd election cycles. Based upon these results we actually see a small, but statistically significant negative effect of adding a new school board member during the even years. This presents some evidence, that often new school board members actually damage the progress of a school district. However, given that the average gain in test scores over 2 years is 24 points, this “damaging,” effect is practically small. One possible reason that we observe a small decline in district performance after 2-years is that it takes time to adapt to the school system, and thus, two years is too small of a period to try and measure performance.

To assess this theory, I also examined the performance of school board members four years after they were elected. A cost to this analysis is that I lose further observations in my sample. Columns four through six show the results of the four year gain. None of these results can be distinguished from zero, although the magnitudes are similar to the 2-year gains. Furthermore, with an average gain of 46 points in my sample, these results are practically equal to zero.

[23] To be complete, I also examined results one and three years after the election with similar results.

A final reason why we might not be observing an effect is that my indicator of a newly elected school board is weak. The indicator is a one if any new members are elected. However, if a single new member is added to a board of five or seven people, they have little power to make changes where a majority is needed to pass any policy. As such, in results not presented here, I also examined cases where all the incumbents running for office were defeated. The results did not differ from what is presented in Table 6.

Certainly when talking about large districts, one could imagine that a two- or four-year time period would be too short to see an effect, but there appears to be little evidence that school board members have much influence on API. In fact, four years after an election, new school board members have a small negative effect. As such, it would make sense if, as we observe in the odd-years, that voters disregard the API scores as a good metric to evaluate school board performance.

1.6. Conclusions

Prior research has shown that the publication of government performance through newspapers, literature, or websites enhances the accountability of elected officials. These findings imply that if the media or government improved reporting on politicians, then government might be more congruent with what voters want. At a time when the national congressional approval rate hovers around 15 to 20 percent, increasing transparency by improving objective information on the effectiveness of elected officials appears to be good public policy to better align congress and their constituents. Similar to other studies on information, I find evidence that, during high turnout years, voters increasingly hold elected officials accountable for their performance. On average, poorer performing incumbents' reelection rates decline by 7 percentage points in the eight years after the implementation of PSAA. We can conclude that, during these high turnout years, information matters.

However, importantly, and central to this study, I find that the effect of information is not uniform across the timing of elections. Based upon my conceptual framework, I predict that those who benefit the most from district policy decisions, teacher and administrators, will turnout in off-cycle elections, while the electorate without children in schools will be less likely. The model then predicts that the off-cycle years will have a different median voter closer to the interests of teachers and administrators. I then find evidence empirically that the interaction between information and voters is different during off-cycle than on-cycle elections. First, I find no statistically significant effect of PSAA on voters during the off-cycle elections. Then I show that there is a 13 percentage point difference, in my preferred specification, between how voters in the off-cycle and on-cycle elections treat PSAA information.

The paper then attempts to uncover one possibility for this effect: school board members have a minimal effect on school performance. I find little evidence that new school board members have a statistically discernible effect on students' test score improvement. There are many reasons why we would observe this effect. First, it takes more than a single term for a new school board member to understand the workings of a district and have a lasting impact. I am unable to test this hypothesis due to the limited size of my panel, but presumably, after a single term, voters would expect to see some results. Second, it could be the case that getting a single new member on a five person board has little effect on the policy decisions made by the school board since it takes a majority of three to make any policy changes. I attempt to test for this by examining the case where two or three new members were added to the school board in an election. While I find no effect for this case either, it is likely there are other unobserved characteristics of an election which I don't account for in my model when none of the incumbents is reelected. Third, as reported in Table 1, incumbents have approximately a 65 percent reelection rate, thus I might not have enough instances in my sample of challengers getting elected to find a statistically significant result. However, the point estimates on my results indicate that during the even-years, where incumbents are held most accountable, there is a practically small negative effect of electing a new school board member. Yet, the most plausible explanation is that test scores might not be the best metric to evaluate school board performance, despite the fact the even-year electorate consistently uses it to hold incumbents accountable.

How then do I interpret the difference between how off- and on-cycle constituents utilize standardized test score information? There are two potential answers. First, prior research (Anzia, 2011; Anzia, 2012; Berry and Gersen, 2011) has shown that teachers' unions wield significant power in off-cycle elections. Teachers' unions have often adopted an adversarial relationship with standardized test scores (Hernandez, 2013). Thus, the disregard of test score information could be because teachers tend to use other metrics, like their union's position, to determine who they vote for during an election, and these effects are stronger on the overall outcomes during off-cycle years. The second explanation is that because mainly teachers, administrators, and active parents turnout for these off-cycle elections, they should have many metrics of performance available to them (e.g., voting records on specific policy issues). If the API is a poor or noisy predictor of school board performance, then it would make sense to disregard this information, especially when compared to voters who know less about school board politics. Unfortunately, my data provides limited ability to test which theory is more applicable in the case of California, and future work must investigate which is more plausible.

References

Anderson, CJ. 2007. "The end of economic voting? Contingency dilemmas and the limits of democratic accountability. *Annual Review of Political Science,* 10: 271-96.

Ashworth, S. 2012. "Electoral Accountability: recent theoretical and empirical work. *Annual Review of Political Science,* 15:183-201.

Anzia, S F. 2011. "Election Timing and the Electoral Influence of Interest Groups." *Journal of Politics* 73 (2): 412-427.

Anzia, S. 2012. "The Election Timing Effect: Evidence from a Policy Intervention in Texas." *Quarterly Journal of Political Science.* 7 (3): 209-248.

Bechtel, MM. and J. Hainmueller. 2011. "How lasting is voter gratitude? An analysis of short- and long-term electoral returns to beneficial policy." *American Journal of Political Science*, 55(4): 851-67.

Banerjee, A, S Kumar, R Pande, and F Su (2010). "Voter Information and Electoral Accountability: Experimental Evidence from Urban India." unpublished manuscript, Harvard University.

Berry, C. and W. Howell. (2007) "Accountability and Local Elections: Rethinking Retrospective Voting." *Journal of American Politics* 69(3): 844-858

Berry, C. and J. Gersen. (2011) "Election Timing and Public Policy." *Quarterly Journal of Political Science* 6(2): 103-135

Bertrand, M., E Duflo, and S. Mullainathan. (2004). "How much should we trust differences-in-differences estimates?" *The Quarterly Journal of Economics* 119(1): 249-275.

Besley, T. 2006. *Principled Agents? The Political Economy of Good Government*. New York: Oxford University Press.

Besley, T. and R. Burgress (2002). The Political Economy of Government Responsiveness: Theory and Evidence from India." *Quarterly Journal of Economics*, 177, 1415-51.

Bjorkman , M. and J. Svensson (2009). "Power to the people: evidence from a randomized field experiment on community-based monitoring in Uganda." *The Quarterly Journal of Economics* 124(2): 735-769.

Bobins, G., L. Camara Fuertes, R. Schwabe (2010). "Does Exposing Corruption Politicians Reduce Corruption?" unpublished manuscript, Massachusetts Institute of Technology.

Chong, A., A. O, D. Karlan, and L. Wantchekon. "Looking Beyond the Incumbent: Effects of Exposing Corruption on Electoral Outcomes." Cambridge, MA: National Bureau of Economic Research, NBER Working Paper No. 17679, December 2011.

Dunne, S., W. Reed, and J. Wilbanks. 1997. "Endogenizing the Median Voter: Public Choice Goes to School." *Public Choice* 93: 99-118

Ferraz, C. and F. Finn (2008). "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics, 123, 703-45.*

Ferraz, C. and F. Finn (2009). "Electoral Accountability and Corruption: Evidence from the Audits of Local Governments." Cambridge, MA: National Bureau of Economic Research, NBER Working Paper No. 14937, April 2009.

Grose A., and B. Oppenheimer. 2007. "The Iraq War, partisanship, and candidate attributes: variation in partisan swing in the 2006 U.S. House elections." *Legislative Studies Quarterly* 32(4): 531-57.

Healy, A. and N. Malhotra. 2009. "Myopic Voters and Natural Disaster Policy." *American Political Science Review,* 103(3): 387-406.

Healy, A. and N. Malhotra. 2013. "Retrospective Voting Reconsidered." *Annual Review of*

Political Science, 16: 285-306.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational evaluation and policy analysis*, *19*(2), 141-164.

Hernandez, J. "Union Chief Recommends Delay in Use of Test Scores" *New York Times.* April 30, 2013.

Hess, F. M. 2002. "School Boards at the Dawn of the 21st Century." National School Boards Association

Karol, D. and E. Miguel. 2007. "Iraq war causalities and the 2004 U.S. presidential elections. *Journal of Politics* 69(3): 633-48.

Kriner, D. and F. Shen. 2007. "Iraq war causalities and the 2006 Senate elections." *Legislative Studies Quarterly* 32: 507-30..

Lim, C., JM Snyder, and D. Stromberg (2010)."Measuring media influence on US state courts." Unpublished manuscript

Moe, T. 2006. "Political Control and the Power of the Agent." *Journal of Law, Economics, and Organization*. 22(1): 1-29

Moe, T. 2009. "Collective Bargaining and the Performance of the Public Schools." *American Journal of Political Science*. 53 (1):156-174

Olken, B. (2007). "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy*, 115(2), 200-49.

Patterson, S. C., & Caldeira, G. A. (1983). Getting out the vote: Participation in gubernatorial elections. *The American Political Science Review*, 675-689.

Reinikka, R. and J. Svensson (2005). "Fighting corruption to improve schooling: Evidence from a newspaper campaign in Uganda." *Journal of the European Economic Association*, 3(2-3), 259-67.

Snyder JM and D. Stromberg (2010). "Press coverage and political accountability." *Journal of Political Economy* 118(2):355–408

Wolfinger, R. E., Rosenstone, S. J., & McIntosh, R. A. (1981). Presidential and congressional voters compared. *American Politics Research*, *9*(2), 245-256.

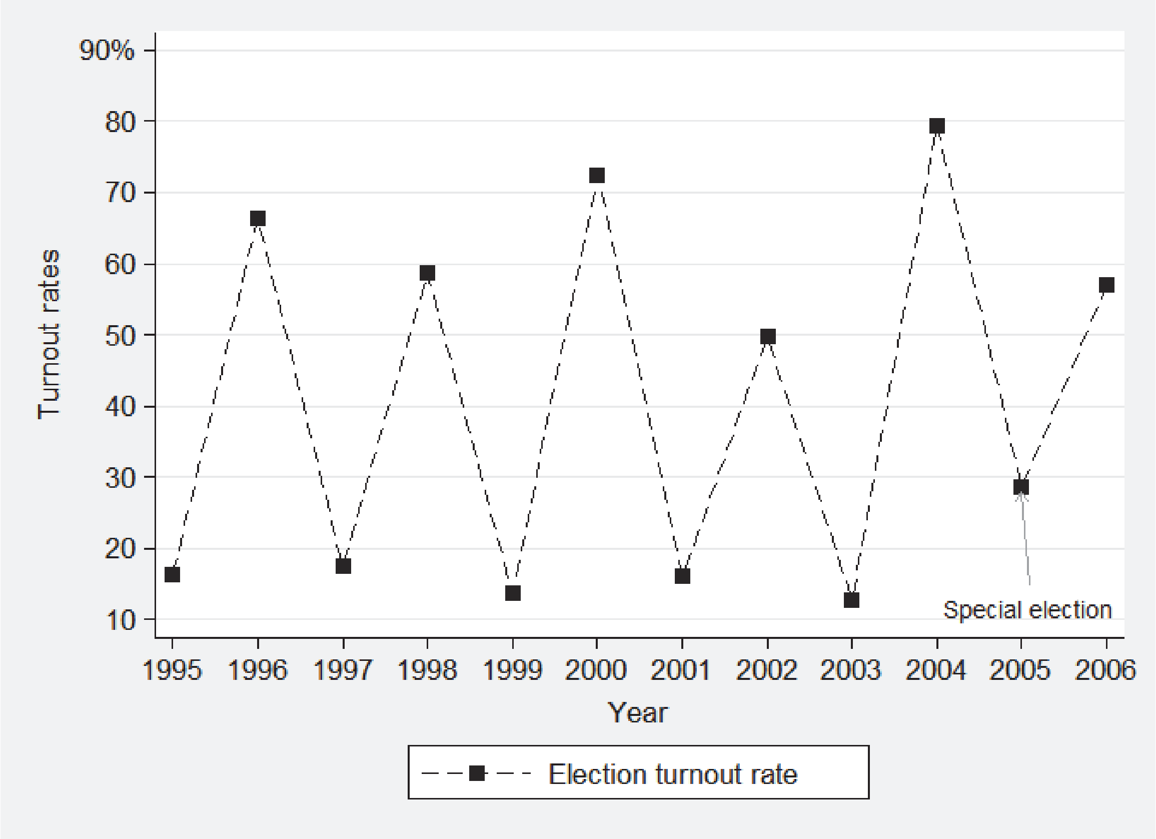

Figure 1.1: Turnout rate among registered voters from 1995-2006

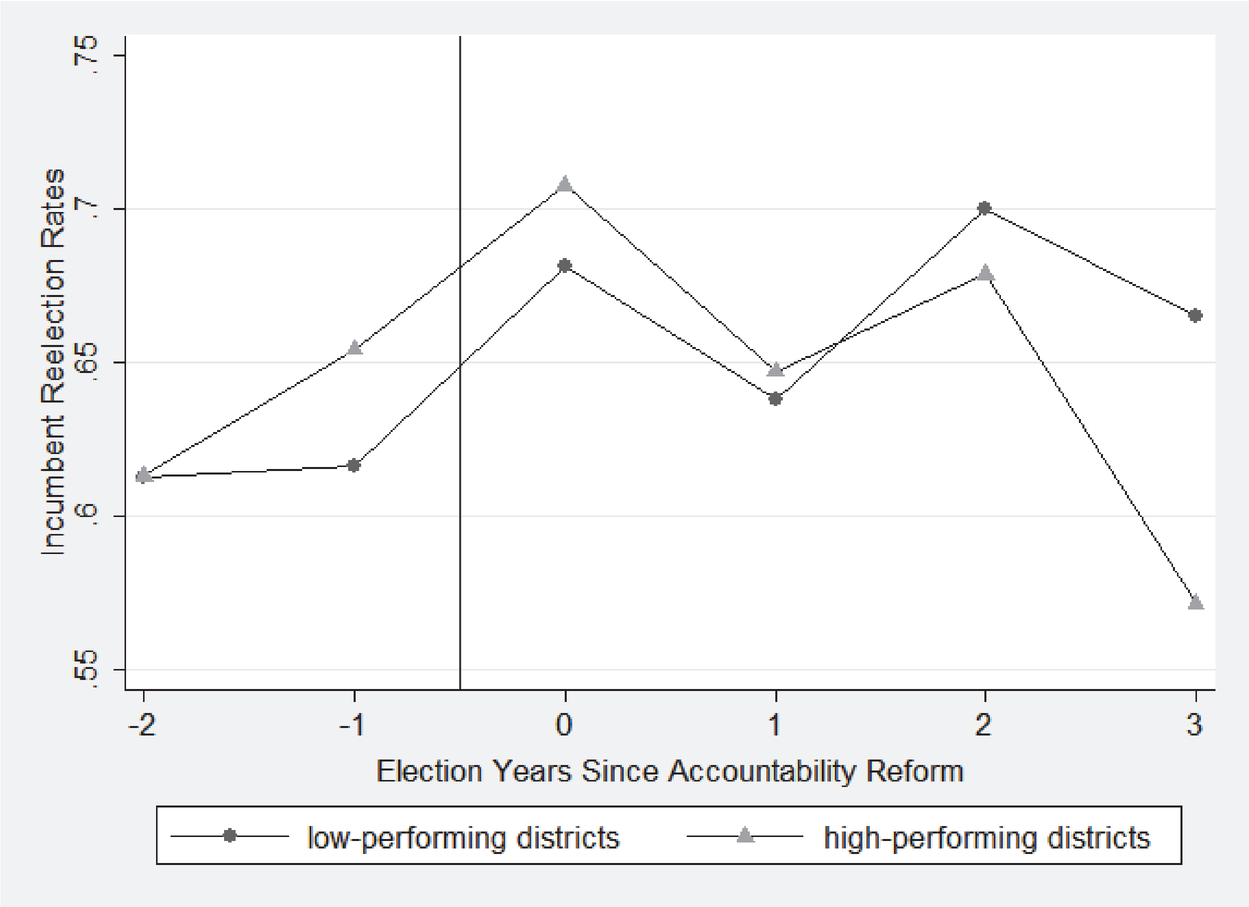

Figure 1.2. Incumbent reelection rates by odd election years since accountability reform

Table 1.2: Difference Estimates of the Effect of Public Schools Accountability Act on Incumbent Reelection Rates During Odd Election Years

		DD		
	(1)	(2)	(3)	(4)
Reelection Rates	0.042	0.050	0.054	0.036
	(0.040)	(0.034)	(0.035)	(0.043)
First Difference	pre-post	pre-post	pre-post	pre-post
Second Difference	low-high API	low-high API	low-high API	low-high API
				excludes 2005
Year Fixed Effects	Y	Y	Y	Y
Controls	N	Y	Y	Y
District Fixed Effects	N	N	Y	Y
Number of districts	297	297	297	297
N	1,782	1,782	1,782	1,782

Note: The dependent variable is the district's incumbent reelection rates for a particular district. The years included are odd years 1995-2005. The affected years are 1999-2005. The demographic controls for students receiving free/reduced price lunch for school, a proxy for poverty levels, and percent Hispanic and black. The N represents district-year observations. Bertrand et al. (2004) serially correlated corrected standard-errors are reported. The standard errors are clustered at the district-level and bootstrapped 400 times. DD = difference in difference; pre-post =post-accountability reform - pre accountability reform. API, the Academic Performance Index, is a composite measure of test scores * indicates significance at the 10% level, ** the 5% level, and *** the 1% level.

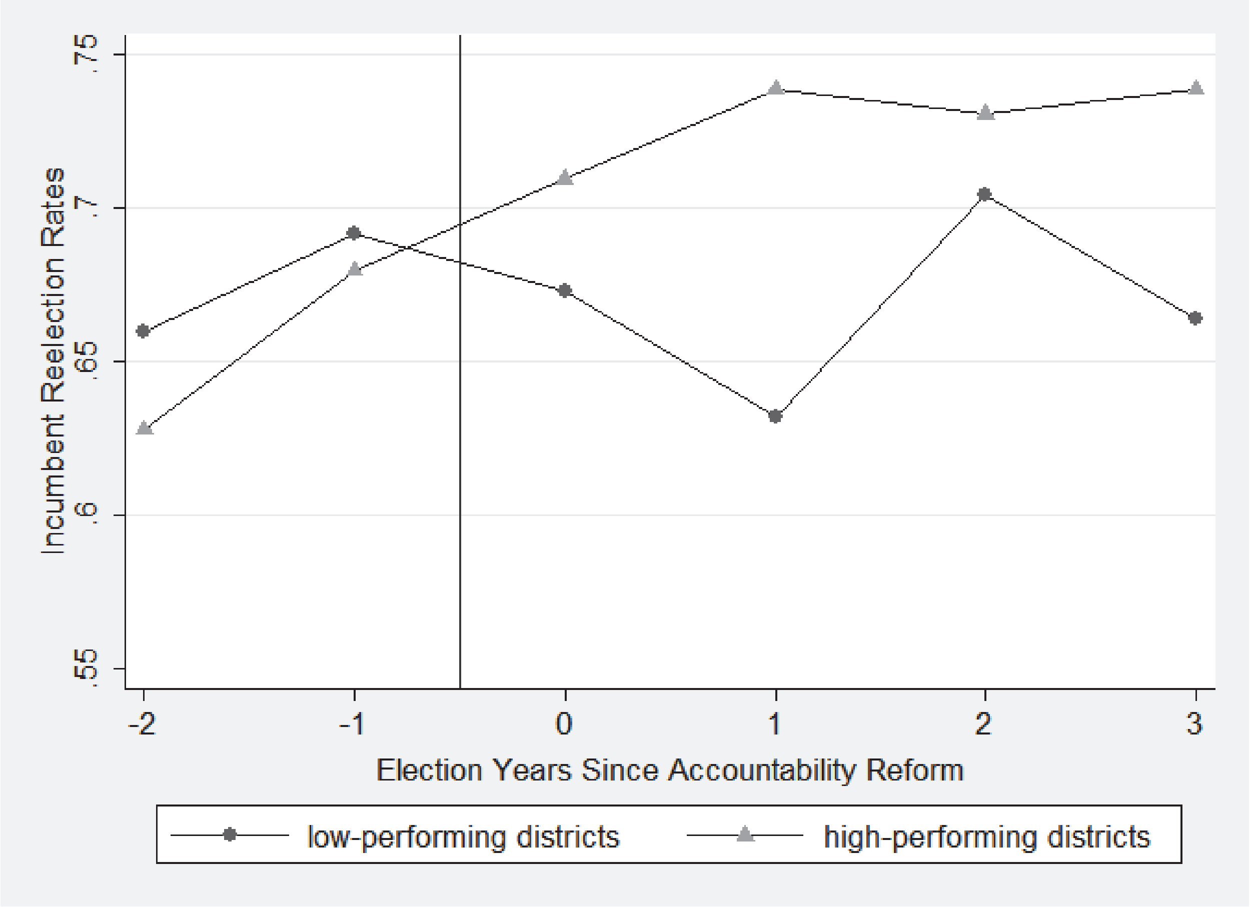

Figure 1.3. Incumbent reelection rates by even election years since accountability reform

Table 1.3: Difference Estimates of the Effect of Public Schools Accountability Act on Incumbent Reelection Rates in Even Years

	Levels		Gains	
	DD	DDD	DD	DDD
	(1)	(3)	(3)	(4)
Reelection Rates	-0.073**	-0.134**	-0.051	-0.123**
	(0.034)	(0.058)	(0.035)	(0.054)
First Difference	pre-post	pre-post	pre-post	pre-post
Second Difference	low-high API	low-high API	low-high API	low-high API
Third Difference		even-odd year election		even-odd year election
Year Fixed Effects	Y	Y	Y	Y
Demographic Controls	Y	Y	Y	Y
District Fixed Effects	Y	Y	Y	Y
Number of districts	452	749	452	749
N	2,712	4,494	2,712	4,494

Note: The dependent variable is the district's incumbent reelection rates for a particular district. The years included are 1995-2006. The affected years are 1999-2006. The demographic controls for students receiving free/reduced price lunch for school, a proxy for poverty levels and percent Hispanic and black. The N represents district-year observations. Bertrand et al. (2004) serially correlated corrected standard-errors are reported. The standard errors are clustered at the district-level and bootstrapped 400 times. DD = difference in difference; DDD = triple difference; pre-post =post-accountability reform - pre accountability reform; API, the Academic Performance Index, is a composite measure test scores. * indicates significance at the 10% level, ** the 5% level, and *** the 1% level.

Table 1.4: Robustness Checks of Estimates of the Effect of Public School Accountability Act on Incumbent Reelection Rates in the even years

	Levels		Gains	
	DD	DDD	DD	DDD
Specifications	(1)	(2)	(3)	(4)
1. Bottom quartile	-0.046*	-0.086**	0.011	-0.004
	(0.029)	(0.043)	(0.021)	(0.029)
2. Continuous treatment (API-SCI)	0.0004*	0.0008	N/A	N/A
	(0.0002)	(0.0006)	N/A	N/A
3. Uncontested elections control	-0.061**	-0.114**	-0.055*	-0.116**
	(0.029)	(0.049)	(0.033)	(0.045)
4. Exclude 2005/2006	-0.071**	-0.102**	-0.049	-0.103**
	(0.030)	(0.051)	(0.031)	(0.049)
First Difference	pre-post	pre-post	pre-post	pre-post
Second Difference	low-high API	low-high API	low-high API	low-high API
Third Difference		even-odd year election		even-odd year election

Note: The dependent variable is the district's incumbent reelection rates for a particular district. The years included are 1995-2006 unless otherwise noted. The affected years are 1999-2006 unless otherwise noted. The demographic controls for students receiving free/reduced price lunch for school, a proxy for poverty levels and percent Hispanic and black. Specification 1 used a dichotomous indicator for whether or not the schools are in the bottom quartile. Specification 2 includes the variable API-SCI in the regression instead of an indicator variable. Specification 3 also includes dummy variable if the election was contested. The N represents district-year observations. Bertrand et al. (2004) serially correlated corrected standard-errors are reported. The standard errors are clustered at the district-level and bootstrapped 400 times. DD = difference in difference; DDD = triple difference; pre-post =post-accountability reform - pre accountability reform; API, the Academic Performance Index, is a composite measure test scores. * indicates significance at the 10% level, ** the 5% level, and *** the 1% level.

Table 1.5: Difference Estimates of the Effect of Public Schools Accountability Act (PSAA) on Incumbent Reelection Rates During the First Two Years of PSAA and for Uncontested Elections

	First Two Years of PSAA	
	DD	DDD
	(1)	(3)
Reelection Rates	-0.051	-0.048
	(0.045)	(0.071)
First Difference	pre-post	pre-post
Second Difference	low-high API	low-high API
Third Difference		even-odd year election
Year Fixed Effects	Y	Y
Demographic Controls	Y	Y
District Fixed Effects	Y	Y
Number of districts	452	749
N	1,356	2,247

Note: The dependent variable is the district's incumbent reelection rates for a particular district. The years included are 1995-2000 for the first two year specification and 1995-2006 for the uncontested specification. The affected years are 1999-2000 for the two year specification and 1999-2006 for the uncontested specification. The demographic controls for students receiving free/reduced price lunch for school, a proxy for poverty levels and percent Hispanic and black. The N represents district-year observations. Bertrand et al. (2004) serially correlated corrected standard-errors are reported. The standard errors are clustered at the district-level and bootstrapped 400 times. DD = difference in difference; DDD = triple difference; pre-post =post-accountability reform - pre accountability reform; API, the Academic Performance Index, is a composite measure test scores. * indicates significance at the 10% level, ** the 5% level, and *** the 1% level.

Table 1.6: Difference Estimates of the Effect of Public Schools Accountability Act on Incumbent Reelection Rates

		Levels			Gains	
	Even	Odd		Even	Odd	
	DD	DD	DDD	DD	DD	DDD
	(1)	(1)	(2)	(3)	(3)	(4)
Rerun Rates	-0.038	0.024	-0.055	0.003	0.027	-0.026
	(0.031)	(0.041)	(0.051)	(0.030)	(0.041)	(0.047)
First Difference	pre-post	pre-post	pre-post	pre-post	pre-post	pre-post
Second Difference	low-high API	low-high API	low-high API	low-high API	low-high API	low-high API
Third Difference			even-odd year election			even-odd year election
Year Fixed Effects	Y	Y	Y	Y	Y	Y
Demographic Controls	Y	Y	Y	Y	Y	Y
District Fixed Effects	Y	Y	Y	Y	Y	Y
Number of districts	477	270	747	477	270	747
N	2,690	1,792	4,482	2,690	1,792	4,482

Note: The dependent variable is the district's incumbent reelection rates for a particular district. The years included are 1995-2006. The affected years are 1999-2006. The demographic controls for students receiving free/reduced price lunch for school, a proxy for poverty levels and percent Hispanic and black. The N represents district-year observations. Bertrand et al. (2004) serially correlated corrected standard-errors are reported. The standard errors are clustered at the district-level and bootstrapped 400 times. DD = difference in difference; DDD = triple difference; pre-post =post-accountability reform - pre accountability reform; API, the Academic Performance Index, is a composite measure test scores. * indicates significance at the 10% level, ** the 5% level, and *** the 1% level.

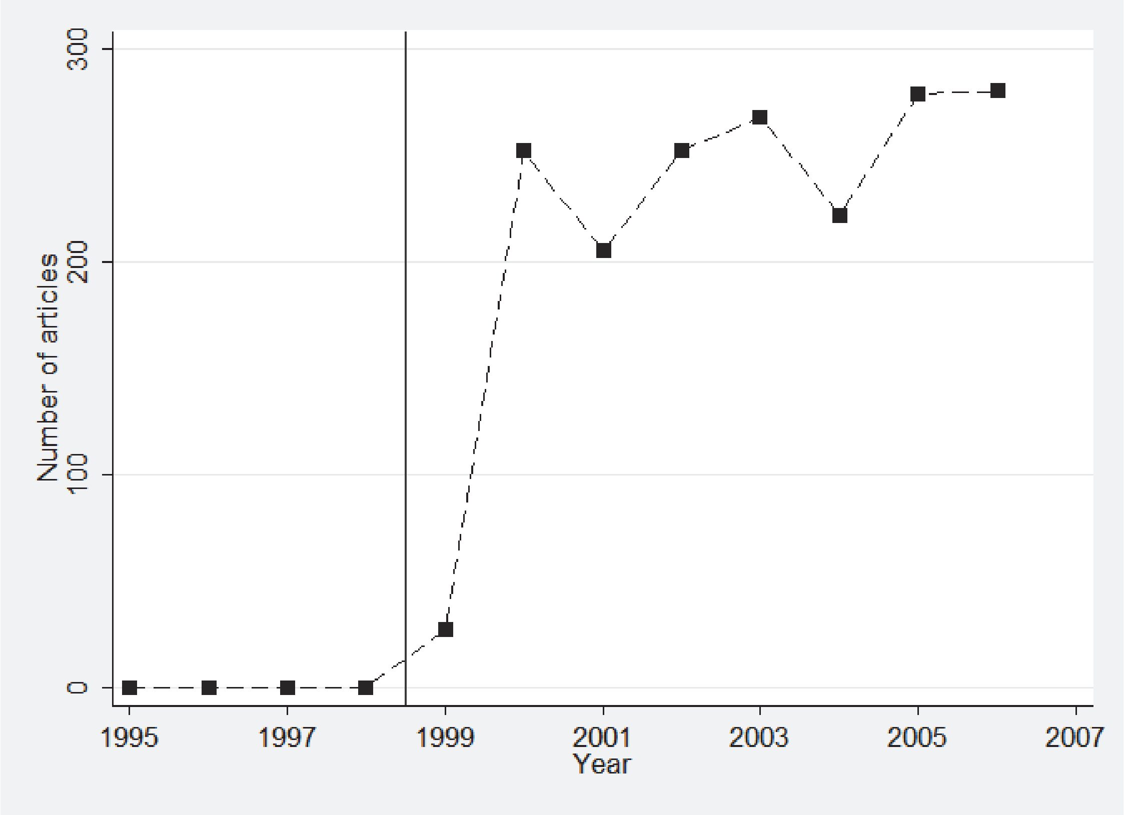

Figure 1.4 – Articles on the "Academic Performance Index" in California

Source: Lexis-Nexis

Table 1.7: Difference Estimates of the Effect of Public Schools Accountability Act (PSAA) on Incumbent Reelection Rates Year-by-Year

Independent Variables	Overall
2 years prior to PSAA *even years*low performers	0.051
	(0.066)
year of PSAA *even years* low performers	0.005
	(0.056)
1 year after PSAA *even years* low performers	-0.098
	(0.063)
2 years after PSAA *even years* low performers	-0.093
	(0.060)
3 years after PSAA *even years* low performers	-0.185**
	(0.059)
Year Fixed Effects	Y
Demographic Controls	Y
District Fixed Effects	Y
Number of districts	749
N	4,494

Note: The dependent variable is the district's incumbent reelection rates for a particular district. The years included are 1995-2006. The affected years are 1999-2006. The demographic controls for students receiving free/reduced price lunch for school, a proxy for poverty levels and percent Hispanic and black. The N represents district-year observations. Bertrand et al. (2004) serially correlated corrected standard-errors are reported. The

standard errors are clustered at the district-level and bootstrapped 400 times. Low performers are schools with similar characteristics, but lower Academic Performance Index scores * indicates significance at the 10% level, ** the 5% level, and *** the 1% level.

Table 1.8. Change in Academic Performance Index after the election of a challenger

	Two year Gains			Four Year Gains		
	all	even	odd	all	even	odd
Challenger Won	-0.013	-0.023*	-0.001	-0.011	-0.019	0.009
	(0.011)	(0.014)	(0.020)	(0.017)	(0.019)	(0.041)
Lag Gains	Y	Y	Y	Y	Y	Y
SCI Controls	Y	Y	Y	Y	Y	Y
N	1931	1299	632	1182	866	316

Note: The dependent variable is the gains in test score after a challenger won. The years included are 2000-2004 for the 2-year gains and 2000-2002 for the 4-year gains. The School Characteristic Index (SCI) controls consist of all the control variables used to predict the SCI (see paper for further explanation). The N represents district-year observations. Standard errors are clustered at the district level. * indicates significance at the 10% level, ** the 5% level, and *** the 1% level.

2. Can recruitment and retention bonuses attract teachers to low performing schools? Evidence from a policy intervention in North Carolina

Abstract

Little evidence exists that bonuses can attract and retain highly effective teachers to low-performing, high-poverty schools, especially at the high school level. This paper investigates whether recruitment, retention, and performance bonuses in coordination with supporting policies affected the distribution of high school teachers within a school district. A companion paper (Scherer, 2014) found that the reform significantly increased student-level achievement. This paper finds that the change in achievement does not appear to be related to comparatively recruiting more effective teachers or removing the least effective teachers. Instead, the progress appears to stem from gains in teacher effectiveness, through a combination of improved leadership and school working conditions as well as performance incentives. These results provide evidence that a combination of high quality leadership, positive school environment, and performance bonuses may be an important policy lever to improve teacher effectiveness in high-poverty, low-performing schools.

Keywords: Recruitment and retention bonuses; Compensating differential

Recently, federal, state, and local government policymakers have focused on methods to improve low-income students' access to effective teachers. This focus is motivated in part by more than two decades of research that empirically shows that teachers are one of the most important inputs into a child's education (see Hanushek and Rivkin, 2006). A second motivation is research using observable teacher characteristics (e.g., experience) shows consistently that lower quality teachers tend to work in lower-performing, high-poverty schools (Clotfelter et al., 2005; Clotfelter et al., 2007; Lankford et al., 2002; and Loeb and Reininger, 2004). In response to these findings, both the federal government (through the Teacher Incentive Fund[24] and Race to the Top[25]) as well as philanthropy (most notably, the Bill and Melinda Gates Foundation[26]) have allocated more than $5 billion to encourage experimentation of states and school districts across the country to measure and incentivize more equitable distribution of effective teachers.

Concurrent with these efforts, the emergence of administrative data sets that track students' test score performance over time and link students to their teachers has drastically changed how a teacher's effectiveness is measured. With these panel datasets, researchers and evaluators can create "value-added" estimates for individual teachers. These models aim to isolate teachers' contribution to student performance. Although this newer metric of teacher effectiveness shows small differences between average teachers in high- and low-poverty schools (Sass et al., 2012; Mansfield, 2014; Buddin, 2011; Max and Glazerman, 2010; Steele et al., 2014; Isenberg et al., 2013), research in this area generally confirms that the distribution of teachers between high- and low-poverty schools is inequitable, but does not comment on how much this gap would change if more high quality teachers were employed in high-poverty schools.

To improve disadvantaged students' access to effective teachers, differential compensation has been put forward a key policy lever. In particular, change advocates aim to provide teachers

[24] Teacher Incentive Fund Fact Sheet, http://www2.ed.gov/programs/teacherincentive/factsheet.html, accessed June 12, 2014

[25] Race to the Top Executive Summary, http://www2.ed.gov/programs/racetothetop/executive-summary.pdf, accessed June 12, 2014

[26] Intensive Partnership for Effective Teaching Press Release, http://www.gatesfoundation.org/Media-Center/Press-Releases/2009/11/Foundation-Commits-$335-Million-to-Promote-Effective-Teaching-and-Raise-Student-Achievement, accessed on June 12, 2014

with recruitment, retention, and performance bonuses to work in low-performing, high-poverty schools. Advocates suggest that changing the salary structure will improve the average applicant quality to a particular school, thereby increasing the likelihood that disadvantaged schools are staffed with equally effective teachers as their more advantaged counterparts. Some very recent empirical evidence shows that teacher recruitment and retention bonuses offered at the elementary and middle school levels can improve the average value-added of teachers educating low-income students (Glazerman et al., 2013; Steele et al., 2014). This research builds on prior work which found that recruitment and retention bonuses offered at lower performing schools attract new teachers and reduce teacher and principal turnover rates (Clotfelter et al., 2008; Steele et al., 2010; Bayonas, 2012).

To gain insight into whether changing the compensation structure can attract and retain effective teachers at high-poverty schools and improve their performance, this paper focuses on the Mission Possible (MP) program. MP was implemented in North Carolina starting in the 2006-07 school year and awarded recruitment, retention, and performance bonuses of up to $14,000 to teachers in high schools serving a high percentage of low-income and low-performing students. The bonuses complemented initiatives to improve school leadership (e.g., principals operating schools with high test scores), professional development, performance accountability, and facility support. The program's theory of change follows a similar logic laid out above where bonuses would improve the applicant pool to higher poverty schools, in turn improving the quality of the teaching labor force, and would improve the performance of incumbent teachers. While the other aspects of the reform may facilitate changes such as improvement of the school's culture, the period of evaluation is insufficient to draw conclusions about their role.

However, it is important to note that any effect I observe would be the combined effect of *all* the reforms rather than only the bonuses.

My analysis uses longitudinal data on North Carolina public school students. First, to understand MP's impact on teachers, I compare trends between MP and non-MP schools. School-level estimates show a narrowing of the gap between MP and non-MP schools with regard to the turnover rate, teacher experience, and credentialing of teachers in the high schools. Next, I estimate teachers' Algebra I value-added score before and after the implementation of the MP program. Using these estimates, I examine whether the effects observed in Scherer (2014) and confirmed in this paper are due to recruiting highly effective new teachers or removing the least effective teachers from the schools. In both cases, I find the schools were unable to attract better teachers or remove poorly performing teachers. Instead, I conclude that the combination of effective leadership, performance bonuses, and possibly improved school working conditions all contributed to these improvements in teacher effectiveness.

This study contributes to the emerging literature on the distribution of effective teachers using value-added models. It is one of the first studies to analyze how the effectiveness (measured using value-added scores) of high school math teachers is distributed across a school district. It is also one of the first studies to examine the change over time in teachers' value-added distribution after a comprehensive district reform. Finally, it is the first study to examine the underlying mechanism that changes the distribution (i.e., whether it comes through hiring or exits from the school) at the high school level.

In section 1 of this paper, I provide a brief overview of the MP program. In section 2, I describe the data set for the difference-in-difference models. In section 3, I present the basic results. A brief conclusion and some policy recommendations are in section 4.

2.1. The structure of the Mission Possible program

Mission Possible (MP) is a multifaceted program to attract and retain effective teachers and administrators in the hard-to-staff schools of Guilford County School District (GCS), North Carolina.[27] GCS is large, diverse, and predominantly low-income. In 2013-14, GCS operated 28 high schools with 23,389 high school students. The student population was 40 percent black, 12 percent Hispanic, and 6 percent Asian. Approximately 60 percent of students received free/reduced price lunch.

The MP program, which was launched in September 2006 and still operates today, provides recruitment, retention, and performance incentives for teachers in select schools (see further description below); staff development; performance accountability; and structural support. The county initially funded the program in 20 of its lowest-performing schools, including nine high schools. Subsequent funding from the federal Teacher Incentive Fund enabled the program to be added to another ten schools for the 2007-08 school year. To select the participating schools, the district calculated a weighted average of each school's percent of students receiving reduced or free meals, Adequate Yearly Progress (AYP) and ABC status[28], and teacher turnover rate.

2.1.1. Recruitment, Retention, and Performance Bonuses

Originally, teachers recruited to MP schools needed to be "highly qualified" according to federal standards.[29] Teachers were screened and interviewed through a process modeled after the

[27] Mission Possible Overview, http://www1.gcsnc.com/depts/mission_possible/background.htm accessed on June 1, 2014.

[28] The ABC's of Public Education is North Carolina's school-level accountability system created during the 1996-97 school year. Adequate Yearly Progress is an indicator of whether particular schools are making progress toward federal accountability goals.

[29] North Carolina State Board of Education Policy Manual, http://sbepolicy.dpi.state.nc.us/Policies/TCP-A-001.asp?Acr=TCP&Cat=A&Pol=001, accessed on June 1, 2014.

Haberman Star Teacher Interview, which focuses on how teachers work with at-risk-youth (Center for Education Compensation Reform, 2008). In later years of the program, teachers were considered for a recruitment award if they had two or more years of above average value-added data.

Table 1 reports the recruitment/retention bonuses provided to teachers and principals. Bonuses ranged from $2,500 to $10,000, with the highest amounts paid to teachers in the hardest-to-staff subject, Algebra I, as well as principals. These bonuses were paid to either attract or retain teachers and administrators in the years after they were recruited.

In addition to the recruitment and retention bonuses, each MP school offered performance bonuses to both teachers and administrators. To identify high performing teachers, the schools used the SAS Institute to calculate individual teacher value-added. Based upon these measures, if a teacher performed one standard error above the mean, he was eligible for a Level I bonus of $2,500. If the teacher performed 1.5 standard errors above the mean, he was eligible for a Level II bonus of $4,000. In order to recognize highly performing administrators, if the school met AYP as defined by the state accountability standards, then the principal was eligible for a $5,000 performance bonus.

Bonuses could represent a significant proportion of a teacher's overall pay. In Guilford County, teacher salaries are determined largely by a statewide schedule that varies by experience level, degree attained, and National Board for Professional Teaching Standards certification status. I was unable to obtain salary schedules for the time of interest and instead used the schedule for the 2013-14 school year, noting that it probably incorporates inflation adjustments. Salaries ranged from just over $35,000 for a teacher with no prior experience and a bachelor's degree to over $66,000 for a teacher with 30 or more years of experience and an advanced degree. A teacher with 15 years of experience and a master's degree earned about $48,000.[30] For this teacher, a combined recruitment, retention, and performance bonus of $14,000 would represent 30 percent of his base salary.

2.1.2. Staff Development

Teachers who received bonuses were required to participate in staff development. In their first year of the program, teachers needed to complete two workshops entitled, *Cooperative Learning* and *Undoing Racism.* Those participating in the second year of the program were to complete workshops on *Differential Instruction* and *Teacher Expectations and Student Achievement*. Teachers had a 13-month window to complete the professional development.

2.1.3. Performance Accountability

The school district also designed a mechanism to remove consistently low performers. In particular, if a teacher had negative value-added measures for more than two years or a principal failed to meet AYP for three years, she was moved to non-MP schools (Center for Education Compensation Reform, 2008).

2.1.4. Structural Support

Finally, the district provided some additional funding to MP schools to reduce the size of mathematics classes and purchase technologies to aid teaching (e.g., graphing calculators) (Center for Education Compensation Reform, 2008).

[30] Guilford County Teacher Salary Schedules, http://schoolcenter.gcsnc.com/education/components/scrapbook/default.php?sectiondetailid=264921&#teacher, accessed on June 1, 2014.

2.2. Data and methods

The North Carolina Education Research Data Center (NCERDC) collects and assembles student-level administrative data on all students in North Carolina. Because the data set allows students to be observed over time, I was able to follow students from 2002 (5 years prior to the intervention) to 2010 (four years post-intervention). I linked the End of Course (e.g., Algebra I) to the End of Grade (e.g., eighth- grade math and reading) exam. I restricted the sample to students who had valid eighth-grade reading and mathematics scores available[31] as well as certain demographic characteristics (e.g., age, race, free/reduced price lunch status).[32] Finally, I separated the sample into MP students and non-MP students within Guilford County.

To link students to their teachers, I matched multiple data files. Algebra I test files have a unique numeric identifier for test proctors, but the proctor was not necessarily the students' class teacher. A separate set of files (the Student Activity Report) identifies teachers at the classroom level, but does not specifically link students to teachers.

To assign teachers to their students, I first linked students' exam proctor code to the teacher code at the classroom level and verified the matches based upon a fit statistic, described in further detail below. This method of matching and then using a fit statistic has been used in other research (Clotfelter, Ladd, and Vigdor, 2007; Mansfield, 2014) to match cases in North Carolina.

To assign students to their likely teachers, I aggregated the students in the Algebra I test file into classrooms by district, school, year, test proctor, and class period. Next, using the Student

[31] This cut the sample size by about 10 percent. While the difference between those with and without eighth-grade math scores is statistically significant, those without the scores at the MP schools were actually higher performers on the Algebra I test. As such, any bias introduced by this sample restriction should reduce the size of the effect.

[32] Note that the administrative data contain other information about the students (e.g., gifted), but I did not include these variables because their quality was inconsistent over time, and it would have further limited the sample. The results did not change when the model was run with more detailed demographics.

Activity Report, I identified all Algebra I classrooms by year, district, school and teacher. I performed a “many to many” match (since there is no common classroom identifier across the two datasets) by year, district, school, and teacher identification variable. Then, similar to Clotfelter, Ladd, and Vigdor (2007), I constructed a fit statistic based upon the sum of the squared deviations for males, whites, and class size. I rejected any matches with a fit statistic greater than 1.5. Using this method, I matched approximately 70 percent of the teachers to their classrooms across all of the years in my study.

Although the match rate was fairly high, a chief concern is whether the matching introduced selection bias into the estimation sample. To check if this was the case, I estimated a t-statistic of the students’ Algebra I test scores between the matched and unmatched teachers during the pre-intervention period. Within Guilford County, the difference was statistically different, with the unmatched students performing worse. However, a further test showed the same result for non-MP schools; the difference between the MP and non-MP schools was not statistically significant at the 0.10 level. Thus, the estimation sample, which is based on the matched data, tends to include higher performing MP and non-MP schools.

Table 2 shows the descriptive statistics for MP students as well as the comparison group in the year prior to the intervention. On average, the MP schools were significantly different from the non-MP schools in their percentage of black students and percentage of students receiving free/reduced price lunch. The MP students performed worse on both the Algebra I test and the eighth-grade math and reading test, although it is only statistically different in the case of Algebra I scores. These descriptive statistics confirm that the MP program targeted schools that were both higher poverty and lower performing than the rest of the district.

My basic estimation strategy involved two steps. I began by estimating the teachers' contribution to students' learning using a "value-added" framework. Similar to Steele et al. (2014), I used a generalized least squares-hierarchical fixed-effects approach described by Borjas and Sueyoshi (1994) and applied to teacher-value added by Aaronson, Barrow, and Sander (2007). They estimated the first stage using the following model:

$$A_{icjt} = \beta_0 + \beta_1 A_{it-1} + \beta_X X_{it} + \beta_Z Z_{ct} + \alpha_{jt} + \varepsilon_{icjt} \qquad (1)$$

A_{icjt} is student achievement for student i in class c assigned to teacher j in year t. Achievement is a function of lagged achievement (A_{it-1}), where I use eighth-grade math and reading scores as an estimate of innate ability and prior learning; observed student-level covariates (X_{it}), such as gender, race/ethnicity, and age; classroom level covariates (Z_{ct}), which include student-level covariates aggregated to the classroom level, as well as class size; teacher value-added (α_{jt}); and unexplained variation in test scores (ε_{icjt}). The Algebra I and math and reading test scores were standardized by subtracting off the state mean and dividing by the state standard deviation in a given year (e.g., z-scores). I estimated a model that included additional lags of test score performance (e.g., seventh-grade math and reading test scores) as well as models excluding the classroom covariates to evaluate the robustness of the results.[33]

Value-added models introduce a number of potential concerns. My models accounts for the fact that test scores (and thus lagged test scores) are measured with error. Similar to Lockwood and McCaffrey (2014) and Briggs and Dominigue (2011), in accounting for the measurement error, I instrument lagged test scores using the lagged test scores from the other subject (e.g., lagged mathematics scores were instrumented by lagged reading scores).[34]

[33] Using these different value-added formulas did not qualitatively change the results.

[34] To be comprehensive, I also instrumented eighth-grade math and reading test scores using seventh-grade math and reading test scores. Using these variables as instruments does not qualitatively change the results presented here.

Student test scores can vary due to random factors like whether the student slept well the night before the exam and whether they get lucky when guessing among multiple-choice answers. These random fluctuations often cancel each other out when averaging over a large number of students. I therefore impose a restriction of a minimum of five students per teacher when reporting estimated teacher effects.

Because of the variability of students within classes from year-to-year, an important consideration in value-added models is whether to use current-year estimates or average the value-added estimates over multiple years. Two studies found that using multiple years of data substantially improved the reliability of the value-added estimates (McCaffery et al., 2009; Schochet and Chiang, 2013). This seems appropriate when value-added estimates are being used for high-stakes personnel decisions. However, I am interested in the year-to-year variation in these estimates. These variations would allow me to assess potential selection during the pre-MP period as well as how the effects change during the post-implementation period (i.e., it is possible that we would not observe effect until several years after the reform is implemented). The reliability concerns are mitigated by the fact that for all of the estimates I am averaging over several teachers (e.g., all the teachers in an MP school in a particular year).

After computing the teacher value-added estimates, I estimated a series of models to better understand the relationship between MP school status in year t and the fitted teacher effects, $\hat{\alpha}_{jt}$ in year t. In particular, to understand the effect of MP on the mean value-added in schools, I first used a simple difference-in-difference model. My empirical approach is based upon the following regression:

$$\hat{\alpha}_{jst} = \beta_0 + \beta_1 Post_t + \beta_2 MP_s + \beta_3 MP_s * Post_t + v_{jst} \qquad (2)$$

As noted above, $\hat{\alpha}_{jst}$ is the fitted teacher effect for teacher j in school s during time period t; MP_s equals one if it is an MP school (the treatment group), and zero if it is a non-MP school; $Post_t$ equals one for the years after the implementation of the program and zero otherwise; and v_{jst} is the residual. Scherer (2014) finds that MP increased achievement in Algebra I test scores. Given those results, mechanically, I expect to find a positive and statistically significant effect on $\hat{\beta}_3$ (i.e., the value-added of teachers in MP schools should increase if student test scores improved). To further refine this model, I change the dependent variable to the standard deviation to understand how the variability of the teachers changes post-MP.

Model 2 also provides a framework for estimating the extent to which any changes in the distribution of teacher effectiveness over time is a function of new hire performance versus improving the effectiveness of existing teachers. To examine whether the distribution of teacher effectiveness is explained by the performance of new hires, I augmented model 2 to estimate the main effect of newly hired status (i.e., first year in any school system), as well as the effect of a three-way interaction among the indicators for MP schools, post-implementation period, and newly hired status, including all underlying two-way interactions.[35]

Finally, I examined whether MP schools improved not only their ability to retain teachers, but to selectively retain the most effective teachers. I addressed this question by predicting $\Pr(exit_{jst})$—the probability that teacher j leaves his or her school s at the end of year t—as a function of whether the school ultimately entered the MP program (MP_s), the teacher's estimated value-added in year t ($\hat{\alpha}_{jst}$), and an indicator of whether it occurred during the post-implementation time period ($Post_t$):

[35] I was also interested in transfers (i.e., teachers new to a particular school, but not new to teaching), however, the majority of new transfers entered MP schools and thus I could not estimate a similar equation to evaluate this effect.

$$\Pr\left(exit_{jst} \middle| MP_s, \hat{\alpha}_{jst}, Post_t\right) = \beta_0 + \beta_1 MP_s + \beta_2 Post_t + \beta_3 Post_t * MP_s +$$

$$\beta_4 \alpha_{jst} + \beta_5 \alpha_{jst} * MP_s + \beta_6 \alpha_{jst} * Post_t + \beta_7 \alpha_{jst} * Post_t * MP_s \qquad (3)$$

The primary parameter of interest is β_7, which captures the probability of leaving an MP as the effectiveness of teachers improves.

I computed the standard errors for the estimates with sandwich estimators clustering on the school. However, a number of papers have raised concerns about the appropriate standard errors for inference in difference-in-difference models when there is a small number of groups (e.g., Bertrand et al., 2004; Donald and Lang, 2007; Cameron et al., 2008; Conley and Taber, 2011). To address this issue, I adapted the method that Abadie et al. (2010) recommend for the Synthetic Control Group (and a form of the model that Bertrand et al., 2004 and Conley and Taber, 2011 recommend for difference-in-difference models). Specifically, I used exact inference with permutation tests based on placebo treatments (Abadie et al., 2010; Bertrand et al., 2004). To do this, I redid the model using a comparison district as a "placebo" treatment site. Doing this for all comparison districts gave a distribution of estimated "treatment" effects for each of the placebo treatments. If the actual estimate of the treatment effect, β_3 in model (2), was large relative to the variability in the distribution of estimated placebo treatment effects, I would conclude that the effect I observed in model (2) was larger than would be expected simply due to random variation. Put another way, the difference would be "statistically significant." Unlike conventional regression statistical inference, this procedure makes no assumptions about outcomes for students being uncorrelated over time or with other students in the same district; it does assume that students' outcomes are uncorrelated across schools in the same district.

2.3. Results

2.3.1. Description of observable high school teacher changes

Prior to examining how the value-added Algebra I teacher distribution changed in MP schools, it is important to understand how the overall high school teacher environment changed during my period of analysis. Figures 1 and 2 show how the percentage of teachers with less than three years of experience[36] and percentage of teachers who were fully licensed,[37] respectively, changed over time for MP and non-MP high schools. Figure 1 shows that the percentage of inexperienced teachers was increasing for both MP and non-MP schools prior to 2007. After the reform, the percentage of inexperienced teachers in both MP and non-MP high schools contracted. I hypothesize that part of the decline resulted from the hiring freeze that many districts were forced to implement in the middle of the Great Recession. However, it is worth noting that prior to MP, MP schools had eight percentage points more inexperienced teachers than non-MP high schools. Post-implementation, the difference decreased to four percentage points. Thus, on average, the experience gap between MP and non-MP schools narrowed after implementation of the MP program.

Figure 2 shows the percentage of high school teachers who were fully licensed over time by MP high school status. While a relatively high percentage of teachers in the MP schools were fully licensed, prior to 2007 the rates had plateaued. There was a significant increase in the

[36] I used three years or less as a category because this is a common metric in the state to evaluate “teacher quality.”

[37] Per the North Carolina Report Card, “when a teacher is referred to as ’fully licensed,’ he/she has met all of the requirements and teaching standards set by the NC State Board of Education for all areas on their license. Teachers who have entered the profession from alternate careers and teachers who have been hired on an ‘emergency’ basis do not have full licenses. In addition, a teacher may be ‘fully licensed’ in one area and have another area in which they are not ‘fully licensed.’ These teachers are classified as not having a full license. Classification is independent of the actual subject/area taught by the teacher. Full licensure is an indication of the level of formal teacher training a teacher has had,” accessed on 6/10/14.

percentage of fully licensed teachers in 2007 (larger than in the non-MP schools), with a subsequent upward trend. Concurrent with the MP reforms, there were state accountability reforms requiring more teachers to be fully licensed, which could explain the upward slope of both MP and non-MP schools after 2007. However, it appears that once again, the gap between the two school types in the percentage of fully licensed teachers narrowed dramatically after 2006.

Finally, Figure 3 compares the overall high school teacher turnover rates between MP and non-MP schools. With the exception of 2003, turnover rates in the MP high schools were high (over 30 percent) and generally on an upward trajectory prior to 2007. Then I observe a dramatic decline in the turnover rates after the implementation of MP. However, non-MP schools also experienced this rapid decline in turnover rates, but the decline is not as steep in the non-MP schools. As such, while some of the decline in turnover might have been due to the Great Recession rather than MP, once again I see a narrowing in the exit rate between the MP and non-MP schools post-reform.

2.3.2. Difference-in-difference model of Mission Possible

I show the distribution of teacher effectiveness for the MP and non-MP schools before and after the implementation of MP in Figure 4. The means between the value-added distributions in the MP and non-MP high schools are different both before and after the reform. The non-MP schools have, on average, more effective teachers than the MP schools prior to the reform, and the reverse after the reform. Lower proportions of teachers are in the bottom tail of effectiveness for MP schools post-reform. The distribution for MP schools is slightly right skewed prior to the intervention, and its skewedness only increases post-reform.

Table 3 shows the basic difference-in-difference estimates for how mean teacher value-added changed after the implementation MP. The first column of Table 3 shows the coefficient estimate on the change in mean value-added in the MP schools compared to the other schools in the district. Cluster-robust standard errors are shown in parentheses and the 90 percent placebo confidence intervals described in the methods section are in brackets. To interpret the placebo confidence intervals, if the point estimate exceeds the confidence intervals, then we call it statistically significant since it is in the top 5 percent of the placebo distribution.

There is some concern that the placebo confidence intervals are too conservative. In particular, the placebo confidence intervals assume that the treated schools have the same variance as the comparison schools. However, if there are significant improvements in the MP schools during the post-implementation period, this may not be a valid assumption. If the variances differ substantially, I would be less able to detect an effect. This is why I show the more classical standard errors as well as the placebo confidence intervals. In addition, I used a less conservative measure of statistical inference, a 90 percent confidence interval, for the placebo measures of statistical significance.

Using the less conservative standard errors, I find a statistically significant 0.35 increase in the effectiveness after the implementation of MP (Table 3) compared to the non-MP schools. This effect is quite large and represents a significant improvement in the quality of the average teacher in MP schools. Furthermore, using the more conservative placebo confidence intervals, I observed that the point estimate of 0.35 exceeds the confidence intervals, making it "statistically significant." Thus, MP appears to have significantly improved the average effectiveness of Algebra I teachers in Guilford County. As noted in the methods section, given that prior research found a significant effect of MP on increasing Algebra I test scores (Scherer, 2014), I would expect to see the mean value-added of teachers improve significantly compared with the non-MP schools.

2.3.3. Difference-in-difference specification check

To check the robustness of the difference-in-difference findings, I analyze a year-by-year comparison between the MP and non-MP schools. I augment equation (2) and substitute leads and lags from the first year of the intervention (e.g., year dummies) for the $Post_t$ variable. Figure 5 plots the estimated leads and lags interacted with the MP_s, running from four years prior to the intervention to four years after the intervention. The omitted category is the year 2002. The y-axis shows the teacher effectiveness in terms of standardized student performance (e.g., z-scores). Figure 5 also plots the two sets of confidence intervals; however the standard errors interpretation will be different. The cluster-robust confidence intervals are shown in black. For an estimate to be statistically significant, these confidence intervals should not include zero. In comparison, using the placebo confidence intervals shown in red, an estimate will be statistically significant if it *exceeds* the confidence range.

The estimates in Figure 5 show no effect in the four years prior to the start of MP, but increased in effect size post-implementation. This pattern demonstrates two things. First, the average teacher's effectiveness prior to the intervention were not statistically different from zero. This point highlights that teacher effectiveness was evenly distributed between non-MP and MP school prior to the intervention, which is consistent with Steele et al.'s (2014) findings. Second, it appears that the bonuses increased teacher effectiveness within the district so that on average, teachers' effectiveness was greater in the MP schools.

Figure 5 shows that on average, teachers were more effective in the MP schools post-intervention. Yet, it is important to understand where the increases in effectiveness occurred (e.g., at the top of effectiveness or the bottom). To investigate this question, I employ a quantile regression and estimate the difference-in-difference at various percentiles of teacher effectiveness. The results are show in Table 4. Once again I show two sets of confidence intervals for each percentile of the analysis. Table 4 indicates that there is broad improvement at all ranges of teacher effectiveness, with a notable exception in the bottom 10 percent using both sets of confidence intervals.

2.3.4. The entry and exit of Algebra I teachers

To better understand the mechanisms through which the improvement in average teacher effectiveness occurs, I analyzed average effectiveness of teachers entering and leaving MP and non-MP schools. Table 5 column (1) shows the interaction coefficient between an indicator of whether the teacher is a new hire, MP_s, and $Post_t$. The coefficient is negative, -0.11, but I cannot distinguish the effect from zero using the cluster-robust standard errors or the placebo confidence intervals. Thus, the improvement in average teacher effectiveness does not appear to be a result of new hires.

Next, I assessed if less effective teachers left MP schools at higher rates than non-MP schools. Column (2) of Table 5 shows $\hat{\beta}_7$ from equation (3). While the size of the coefficient is large and negative, I cannot distinguish it from zero. Thus, it does not appear that MP schools were more effective at removing low-performing teachers than non-MP schools. Based upon these two results, it appears that the increase in teacher effectiveness that I observed must be due to an increase in the productivity of existing teachers. Although I am unable to determine if these productivity increases stem from improved school leadership, professional development, facility support, or the performance incentives offered to the teachers, the findings indicate that MP, taken as a set of changes led to improved teacher performance.

2.4. Conclusion and policy implications

Nationwide, a large gap between the academic performance of low- and high-poverty students persists (Reardon, 2011). Prior research has shown that teachers are one of the most important school factors affecting student achievement, so policy levers that attempt to improve the average teacher quality might be one of the best ways to mitigate the problem. This study focused on a recruitment, retention, and performance bonuses program in North Carolina that was used to attract high-quality teachers to lower performing, higher poverty schools and encourage them to stay. These efforts were motivated by prior research in North Carolina that found differences in the mean performance of low- and high-poverty schools (Sass et al., 2012) as well as mounting efforts by the federal government (MP received a Teacher Incentive Fund Grant in its second year of operation) and the philanthropic community to encourage equal access to effective teachers through differential compensation. However, questions remain about the effectiveness of bonuses in a sector that has typically differentiated jobs through non-pecuniary compensation (e.g., geographic location, working conditions).

This study found that bonuses, along with the other components of the MP program (e.g., more effective school leadership) can improve teachers' value-added. These effects were persistent over the four years post-implementation. This result using a narrower specification (i.e., teacher value-added) replicates the findings in Scherer (2014) that MP schools outperformed other similar schools in the state. Further, this paper found that the improved performance of these high schools does not appear to be due to differential recruitment or removal of teachers, using value-added scores as a metric. Prior work noted turnover rates declined with bonuses, but my work indicates that while overall turnover rate declined, they were unable to selectively retain the most effective teachers. Instead, the improvement likely stemmed from the combination of retention bonuses leading to relatively lower turnover rates, performance bonuses, effective leadership, and possibly improved working conditions within the school. These results suggest that a set of policy actions such as those embodied by MP can improve the performance of incumbent teachers. Although MP did not appear to improve the recruiting of high quality teacher and the separation of lower quality teacher, perhaps stronger incentives would do so; that is a matter for future research.

Because of the difficulty in parsing different aspects of the reform, several important policy questions remain about the underlying mechanisms of this achievement improvement. First, and importantly, what role did the performance incentives play for teachers and what would happen if they were removed in subsequent years? If the performance incentives helped improve teaching practices, arguably causing a permanent shift in effectiveness, then policy-makers could safely remove the performance bonuses after several years without a drop in student achievement. However, if the performance bonuses induced an effort to improve, removing them would result in a subsequent decline in achievement. Existing data could be used to test this hypothesis by examining teachers who received bonuses at MP schools and subsequently transferred to non-MP schools. However, these types of changes occurred too infrequently during the post-MP period of my study to test this hypothesis.

Moreover, if improved achievement resulted because performance bonuses increased teachers' efforts, as the gap between MP and non-MP schools decreased, high value-added scores would be more difficult to obtain (i.e., incoming students would be higher performers so obtaining large gains would be more difficult to achieve). Thus, in order to maintain levels of effort, would higher levels of compensation be needed? In light of prior research showing that performance bonuses alone (individual and team) do not appear to be effective at increasing effort (Springer et al., 2010; Springer et al., 2012), there is some comfort that these improvements in achievement are not solely due to enhanced effort. However, subsequent research should test this hypothesis.

Finally, while Scherer (2014) showed that MP improved student achievement for high school Algebra I, the analysis was unable to isolate the cost of MP for Algebra I teachers (e.g., the amount of recruitment, retention and performance bonuses allocated to teachers and principals as well as the cost of professional development). Similar to Race to the Top,[38] systems where all compete but not all win could be a very cost-effective way to improve performance. However, without more precise estimates of cost, I cannot compare MP to other types of reforms. Future research needs to address this gap in the literature.

[38] Race to the Top Executive Summary, http://www2.ed.gov/programs/racetothetop/executive-summary.pdf, accessed June 12, 2014.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490).

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. *The Quarterly Journal of Economics*, *119*(1), 249-275.

Borjas, G. J., & Sueyoshi, G. T. (1994). A two-stage estimator for probit models with structural group effects. *Journal of Econometrics*, *64*(1), 165-182.

Briggs, D., & Domingue, B. (2011). Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the" Los Angeles Times". *National Education Policy Center*.

Buddin, R. (2011). Measuring Teacher Effectiveness at Improving Student Achievement in Los Angeles Elementary Schools. *Santa Monica, Calif.: RAND Corporation, unpublished working paper*.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414-427.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education review*, *24*(4), 377-392.

Clotfelter, C., Ladd, H. F., Vigdor, J., & Wheeler, J. (2006). High-poverty schools and the distribution of teachers and principals. *NCL Rev.*, *85*, 1345.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682.

Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, *92*(5), 1352-1370.

Conley, T. G., & Taber, C. R. (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, *93*(1), 113-125.

Donald, S. G., & Lang, K. (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, *89*(2), 221-233.

Glazerman, S., & Max, J. (2011). Do Low-Income Students Have Equal Access to the Highest-Performing Teachers? (No. 6955). Mathematica Policy Research.

Glazerman, S., Protik, A., Teh, B. R., Bruch, J., & Seftor, N. (2012). Moving High-Performing Teachers: Implementation of Transfer Incentives in Seven Districts (No. 7412). Mathematica Policy Research.

Glazerman, S., Protik, A., Teh, B. R., Bruch, J., & Max, J. (2013). Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. NCEE 2014-4003. *National Center for Education Evaluation and Regional Assistance*.

Goldhaber, D., & Hansen, M. (2013). Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance. *Economica*, *80*(319), 589-612.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education*, *2*, 1051-1078.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37-62.

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for Test Score Measurement Error in ANCOVA Models for Estimating Treatment Effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 22-52.

Loeb, S., & Reininger, M. (2004). Public Policy and Teacher Labor Markets. What We Know and Why It Matters. *Education Policy Center*.

Mansfield, R. K. (2014) "Teacher Quality and Student Inequality" Unpublished Working Paper.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572-606.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither opportunity*, 91-116.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics*, *72*(2), 104-122.

Schochet, P. Z., and Chiang, H. S. (2013). What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models?. *Journal of Educational and Behavioral Statistics*, *38*(2), 142-171.

Steele, J. L., Baird, M., Engberg, J. & Hunter, G. (2014). Trends in the distribution of teacher effectiveness in the Intensive Partnership for Effective Teaching. *Unpublished manuscript*.

Steele, J. L., Murnane, R. J., & Willett, J. B. (2010). Do financial incentives help low-performing schools attract and keep academically talented teachers? Evidence from California. *Journal of Policy Analysis and Management*, *29*(3), 451-478.

Table 2.1: Incentive Structure for the Original Mission Possible Schools

	Principals	Algebra I	Mathematics Teacher (with math degree)[1]	Mathematics Teacher (without math degree)[1]	English I
Recruitment and retention bonus	$10,000	$10,000	$9,000	$2,500	$2,500
Performance bonus					
Level 1	N/A	$2,500	$2,500	$2,500	$2,500
Level 2	N/A	$4,000	$4,000	$4,000	$4,000
School meets AYP	$5,000	N/A	N/A	N/A	N/A

[1] Individuals also needed 24 hours of training in the content area that they would teach

Note: Conditions for Level 1 and Level 2 bonuses were based upon the SAS Institutes' teacher value-added. Teachers one standard error above the mean qualified for a level 1 bonus, while those 1.5 standard errors above the mean qualified for level 2 performance bonus. AYP stands for Adequate Yearly Progress as defined by the North Carolina state standards.

Table 2.2: Student characteristics in analytic sample by school type

Student Characteristics	Non-MP High Schools	MP High Schools
Student Characteristics		
Black	45%	66%
Hispanic	9%	11%
Asian	5%	5%
Other	4%	4%
Free/Reduced price lunch	57%	74%
Female	50%	47%
Age	15.27	15.36
Student performance		
Algebra I z-score	-0.62	-0.81
8th grade math z-score	-0.70	-0.81
8th grade reading z-score	-0.61	-0.72

Note: All numbers are for 2006. Algebra I and 8th grade math and reading scores are standardized by subtracting the state mean and dividing by the state standard deviation within grade and year. Algebra I scales score, black and free/reduced price lunch are statistically different at the .01 level.

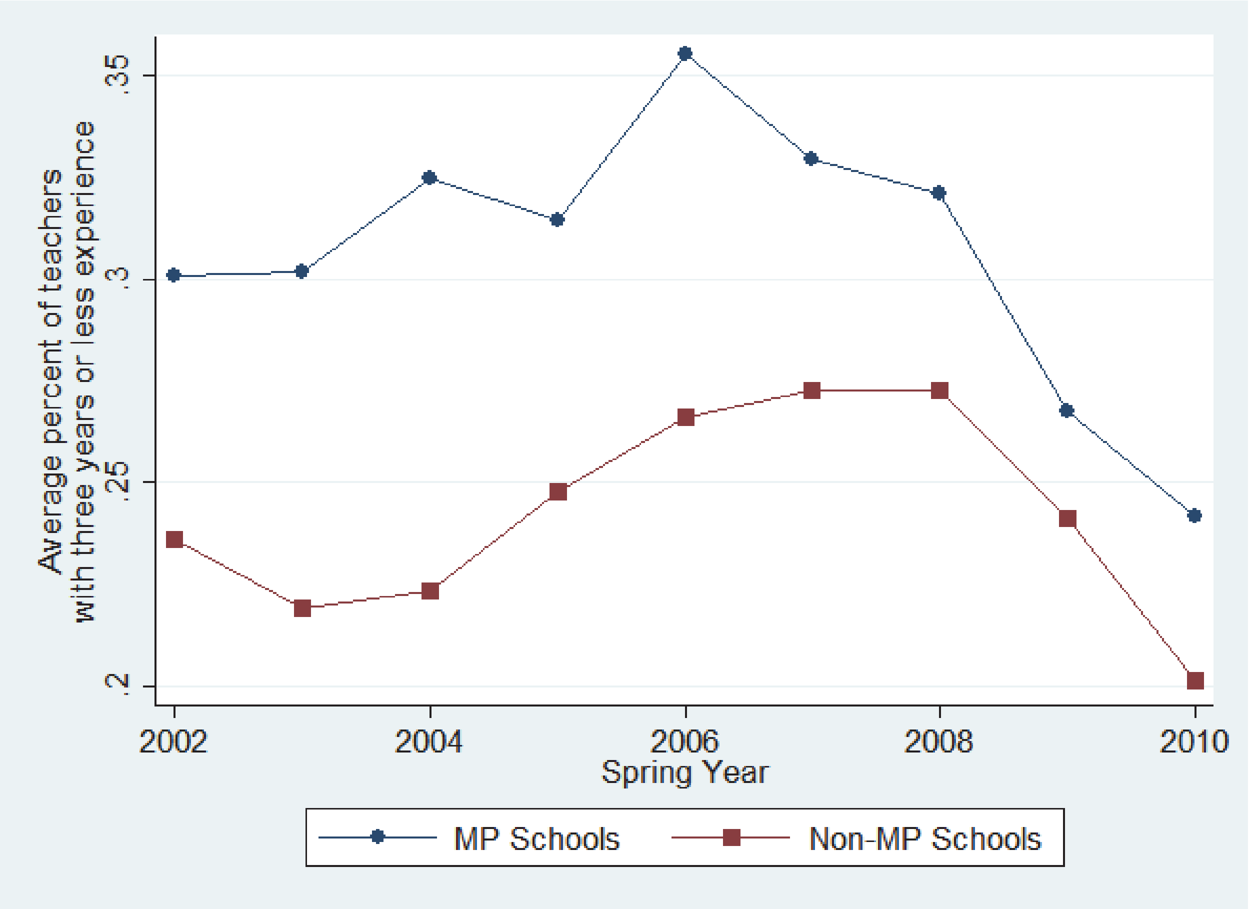

Figure 2.1: Average percent of teachers with three years or less experience over time by whether the school is a Mission Possible school.

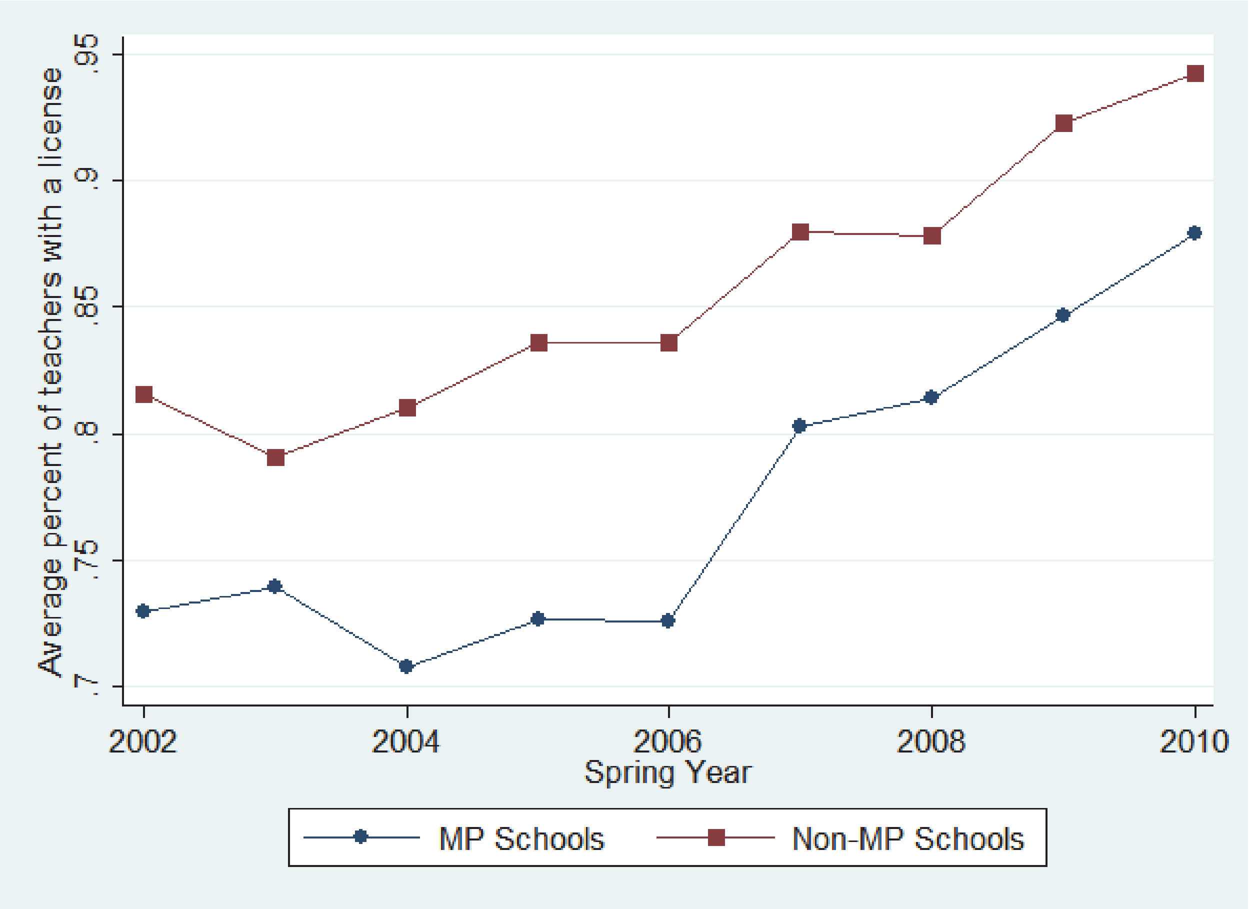

Figure 2.2: Average percent of teachers fully licensed over time by whether the school is a Mission Possible school.

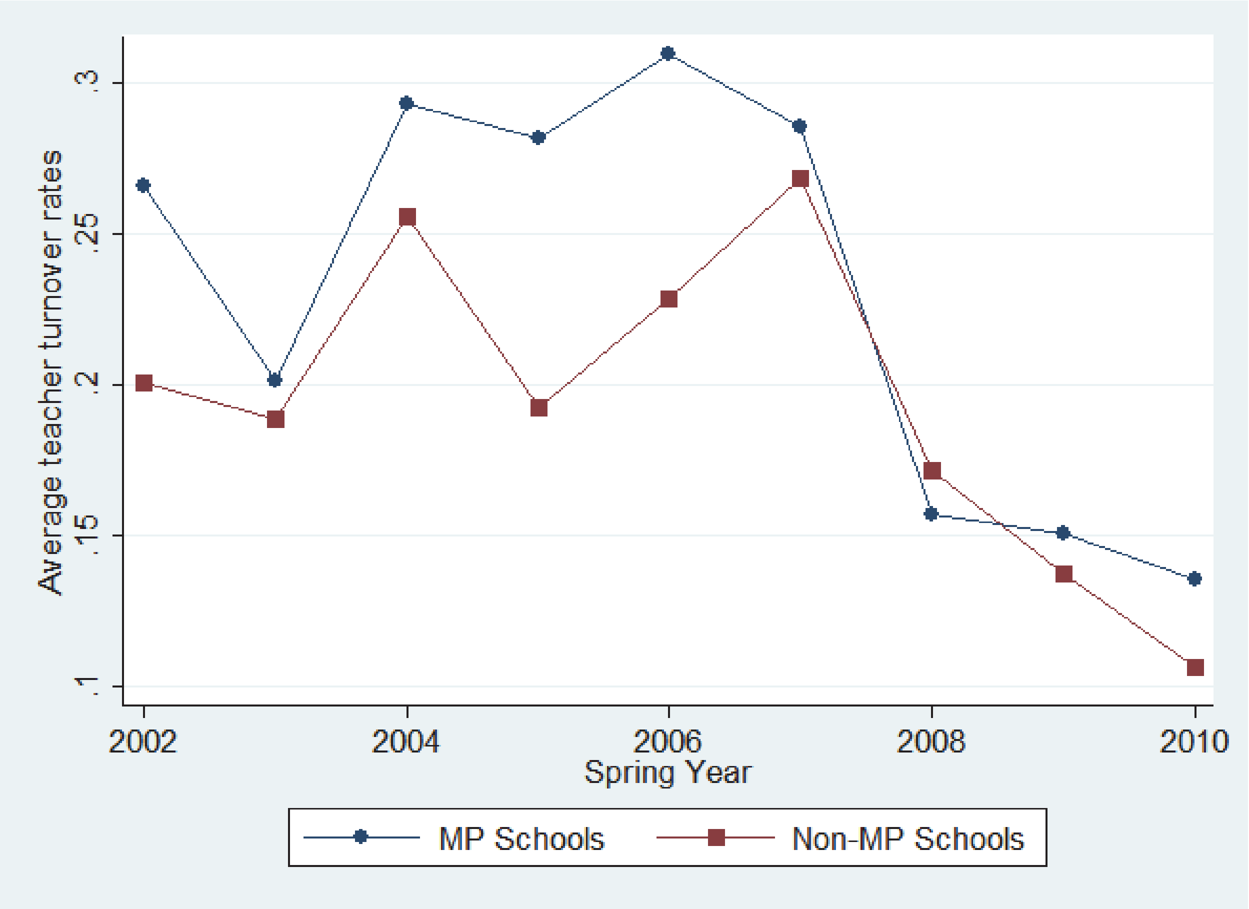

Figure 2.3: Average turnover rates over time by whether the school is a Mission Possible school.

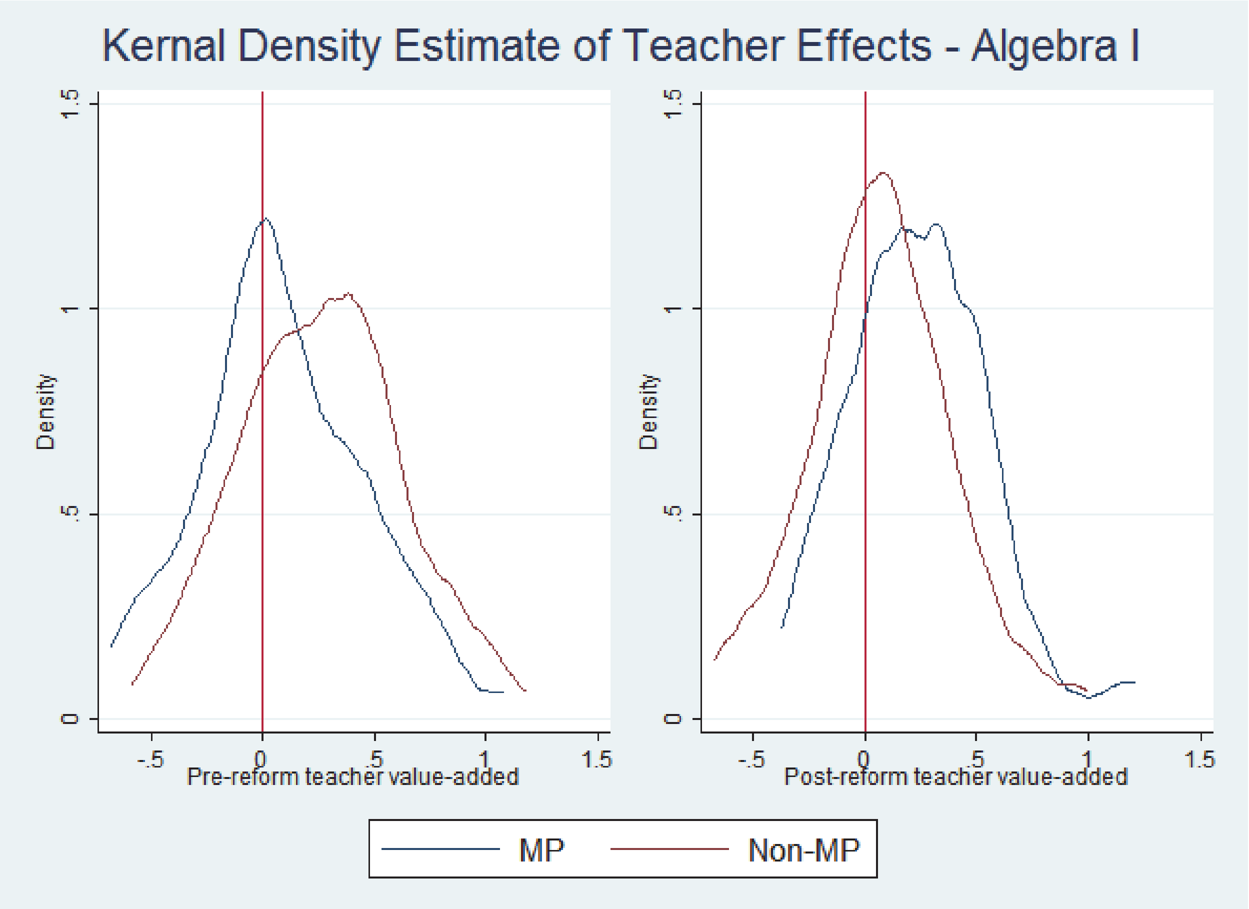

Figure 2.4: Distribution of teacher value-added for MP and non-MP schools, before and after the implementation of MP. The kernel is the Epanechnikov kernel with a bandwidth equal to 0.10.

Table 2.3 Average and standard deviation improvements in teacher value-added at Mission Possible schools

	Means	Standard Deviation
MP*Post	0.35*	0.03*
Standard error	(0.09)	(0.01)
Placebo confidence intervals	[-0.52, 0.09]	[-0.06, 0.02]
N	593	593

Note: Difference-in-difference estimates from a regression of teacher value added. Standard errors are cluster-robust sandwich estimator. Placebo confidence intervals calculated by redoing the difference-in-difference estimator by assigning a comparison school as a placebo "treatment" site. We redo this for all high schools within Guilford county that existed prior to 2007. We then identify the confidence interval by examining the percentiles within the distribution. Estimates that exceed the 90 percent placebo confidence intervals are in the top five percent of the placebo distribution.

*statistically significant using the placebo confidence intervals

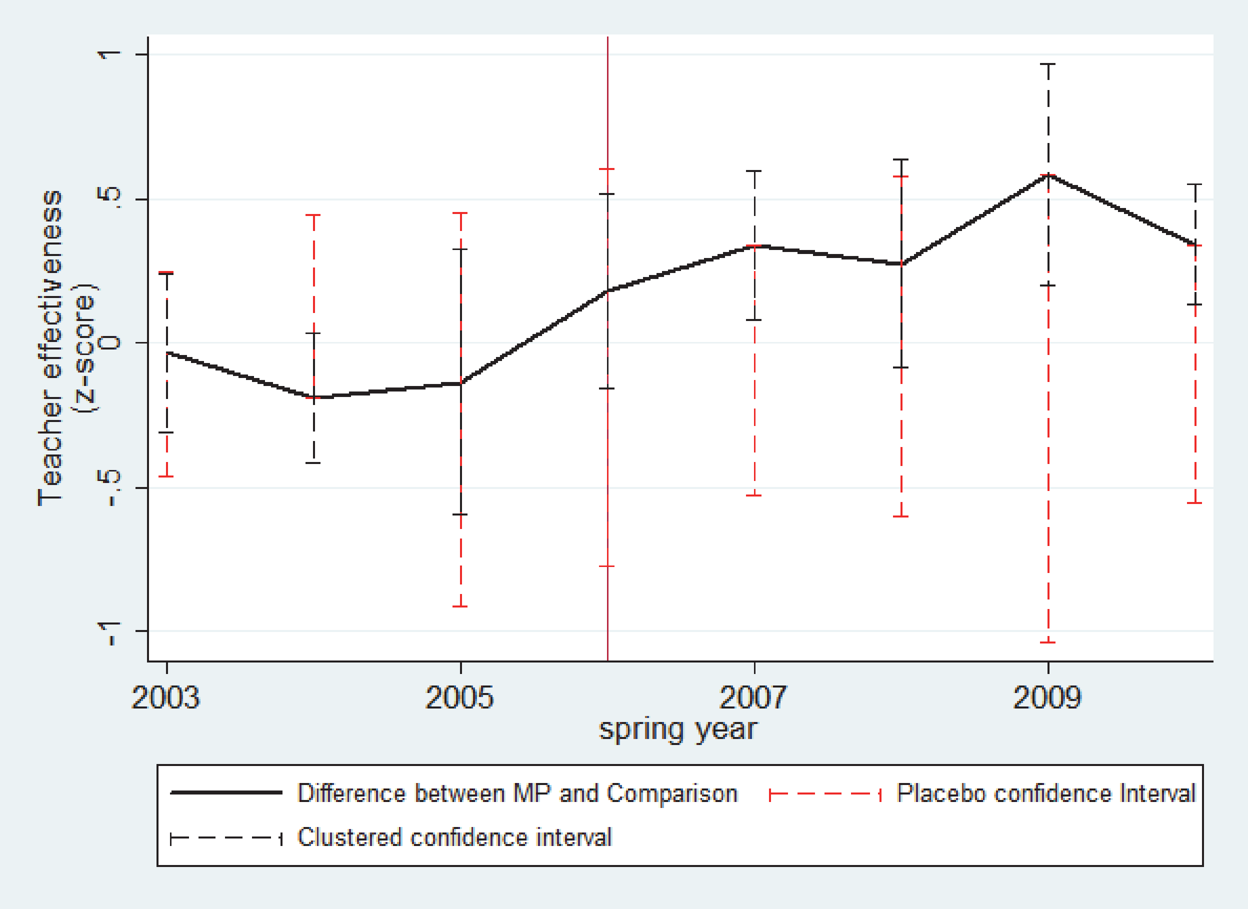

Figure 2.5 Difference-in-difference estimates of MP on standardized teacher effectiveness by year with clustered standard errors and placebo confidence intervals.

Table 2.4 Difference-in-difference estimate at various teacher effectiveness percentiles

Teacher performance percentile	MP*Post	Standard error	Placebo confidence intervals
5	0.25	0.18	[-0.47, 0.25]
10	0.16	0.12	[-0.41, 0.18]
15	0.26*	0.12	[-0.48, 0.20]
20	0.28*	0.10	[-0.44, 0.13]
25	0.31*	0.10	[-0.63, 0.19]
30	0.41*	0.11	[-0.64, 0.18]
35	0.43*	0.10	[-0.61, 0.16]
40	0.36*	0.11	[-0.64, 0.09]
45	0.31*	0.10	[-0.67, 0.15]
50	0.35*	0.11	[-0.67, 0.12]
55	0.28*	0.11	[-0.74, 0.11]
60	0.30*	0.11	[-0.75, 0.12]
65	0.40*	0.11	[-0.76, 0.02]
70	0.40*	0.10	[-0.73, 0.20]
75	0.39*	0.09	[-0.51, 0.14]
80	0.36*	0.10	[-0.57, 0.06]
85	0.39*	0.12	[-0.59, 0.17]
90	0.33*	0.14	[-0.47, 0.30]
95	0.33*	0.17	[-0.85, 0.01]

Note: Difference-in-difference estimates from a quantile regression of teacher value added. Standard errors are cluster-robust sandwich estimator. Placebo confidence intervals calculated by redoing the difference-in-difference estimator by assigning a comparison school as a placebo "treatment" site. We redo this for all high schools within Guilford county that existed prior to 2007. We then identify the confidence interval by examining the percentiles within the distribution. Note that there are only 8 other high schools in Guilford county that did not receive the treatment, and hence show the range of 8 schools in the sample (e.g., the confidence intervals represent the lower 8 schools in the sample).
*statistically significant using placebo confidence intervals

Table 2.5 The value-added of teachers entering and exiting MP schools, compared to non-MP schools.

	New Hires	Exits
Interaction term	-0.11	-0.36
Standard error	(0.19)	(0.23)
Placebo confidence intervals	[-0.98, 1.37]	[-0.72,1.04]
N	593	536

Note: Difference-in-difference estimates from a regression of teacher value added in column (1) and teacher exits in column (2). The interaction term in column (1) is between, an indicator for MP status, Post, and an indicator if the teacher is a new hire. The interaction term in column (2) is between an indicator for MP status, Post, and the continuous measure of teacher effectiveness, value-added. There are fewer observations for the exit interaction because I excluded 2010 from the analysis (see text for further description why). Standard errors are cluster-robust sandwich estimator. Placebo confidence intervals calculated by redoing the difference-in-difference estimator by assigning a comparison school as a placebo "treatment" site. We redo this for all schools within Guilford county. We then identify the confidence interval by examining the percentiles within the distribution. Note that there are only 9 other high schools in Guilford county that did not receive the treatment, and hence we use 80% confidence intervals.

3. Can principal and teacher bonuses improve student achievement? Evidence from a policy intervention in North Carolina

Abstract

Recent experimental evidence on teacher bonuses has cast doubt on their ability to improve student performance. An innovative program in North Carolina, however, provides encouraging new evidence that targeted financial awards for teachers and principals may have a strong positive effect on academic achievement. During the 2006-07 school year, a North Carolina school district implemented an innovative comprehensive recruitment, retention, and performance bonus program for teachers and principals that differed significantly from prior experiments. Using longitudinal data on the students in these schools, I identify the impact of this program on students' math scores by estimating a difference-in-difference model. Results suggest that while the program was no effective during the first three years of implementation, in the fourth year it shifted student performance by 0.28 standard deviations. These results provide evidence that targeted bonuses for teachers and principals could be used to improve academic performance in schools with large proportions of low-income students.

Keywords: Recruitment and retention bonuses; Compensating differential

A positive work environment in coordination with the appropriate level of compensation are important factors in determining whether a high-quality employee remains at the same firm over time. Yet, as a starting point, many urban high-poverty and low-performing schools are uniformly difficult places to work. Teacher groups have raised concerns about the quality of the teaching environment in these settings (Marsh et al., 2011). Possibly because of the challenging environments, these schools are often plagued by frequent turnover in leadership and teaching staff. These problems drive down the applicant pool for open positions, resulting in little to no competition for positions or positions remaining unfilled. While higher compensation is often used to offset a poor work environment in other sectors, the teacher and principal labor market offers salary differentials only based upon degree level and years of experience. As a result, school district employees are highly motivated by non-pecuniary incentives, like better working conditions, driving many competitive teachers to higher performing schools (Greenberg and McCall, 1974; Lankford et al., 2002; Hanushek et al., 2004; Clotfelter et al., 2006; 2007). These conditions might create a skewed distribution of effective teachers and principals away from low-poverty schools. The driving concern is that if high-poverty schools are disproportionally populated by lower quality teachers, the educational opportunities provided to students are not equal.

Whether and how the structure of the principal and teacher labor market affects the distribution of high-quality educators across schools remains a debated empirical question. Research using observable characteristics (e.g., experience, licensure, teacher licensure test scores, and competitiveness of undergraduate institutions) shows consistently that lower quality teachers (Clotfelter et al., 2005; Clotfelter et al., 2007; Lankford et al., 2002; and Loeb and Reininger, 2004) and principals (Loeb et al., 2010; Branch et al., 2009) tend to work in lower-performing, high-poverty schools. However, more recent studies using value-added estimates (an attempt to measure an individual's contribution to student learning) shows mixed evidence on whether high-poverty students receive more effective teachers on average (Sass et al., 2012; Mansfield, 2014; Buddin, 2011; Steele et al., 2014; Isenberg et al., 2013).

While the empirical evidence is mixed, policy-makers are trying to increase the supply of high-quality teachers and principals in lower-achieving, high-poverty schools by providing teachers and principals incentives to work in these schools. By providing recruitment and retention incentives, for instance, districts can compensate teachers and principals for more difficult work environments. Some recent empirical evidence shows that teacher recruitment and retention bonuses offered at the elementary, middle, and high school levels can improve the average value-added of teachers within a school with a high proportion of low-income students (Glazerman et al., 2013; Scherer 2014; Steele et al., 2014). This research builds on prior work which found that recruitment and retention bonuses offered at lower performing schools attract new teachers and reduce teacher and principal turnover rates (Clotfelter et al., 2008; Steele et al., 2010; Bayonas, 2012).

Importantly, improving principal quality and stability can have a broader effect on the work environment of high-poverty schools. For example, based upon a cross-sectional analysis of the North Carolina working-conditions survey, the most important factor in teacher reported and actual departure rates relates to perceived quality of the leadership (Ladd, 2011). Another study of first-year teachers in New York City finds that teachers' perceptions of the school administration is the most important working-condition factor on their retention decision (Boyd et al., 2011). Finally, a longitudinal study also using the North Carolina working-conditions survey data shows that improving the quality of the principal has an influence on every other aspect of the school environment (Burkhauser, 2014).

All of these efforts—attracting and retaining talent as well as improving working conditions of staff in low-performing schools—are intermediate steps to improve the academic achievement of students. This paper investigates whether changing the compensation structure at high-poverty schools can have an effect on the final outcome of interests: student achievement. For this investigation, I examine the effectiveness of the Mission Possible (MP) program. Implemented in North Carolina starting in the 2006-07 school year, MP awarded recruitment, retention, and performance bonuses of up to $15,000 for principals and $14,000 for teachers in high schools serving large proportions of low-income and low-performing students. Principals and teachers were also provided professional development and facility support. The program's theory of change follows a similar logic laid out above where recruitment bonuses would improve the applicant pool, in turn improving the quality of the leadership and the teaching labor force. Annual retention and performance bonuses would increase retention of high-quality staff. Improvement to principal leadership would also improve teacher retention and the overall working conditions of the school. Ultimately, all of these improvements would improve student performance through improved teaching quality. Preliminary analysis of MP suggests that some of the intermediate steps were successful (Scherer, 2014). This research focuses on whether those intermediate steps led to the desired outcome of improved student achievement.

My analysis uses longitudinal data on North Carolina public school students to estimate the effect of MP on ninth grade Algebra I test scores. I employ a difference-in-difference model to compare the MP schools with the remainder of students in North Carolina. In order to account for potential differences between students in the MP schools and the comparison group, I also utilize propensity score weights. The results suggest that the combined set of bonuses in coordination with the other components of the program had large and significant effects on the students in the MP schools. In particular, there was an average improvement of 0.28 standard deviations in the fourth post-intervention year. These findings compliment companion research on MP that providing individual retention and performance incentive compensations for teachers can improve the value-added of teachers within hard-to-staff schools (Scherer, 2014). Taken together, these papers demonstrate to policy-makers that well designed incentive compensation strategies can be effective levers for student improvements.

In section 1 of this paper, I provide a brief overview of the MP program. In section 2, I describe the data set I construct for the difference-in-difference and year-by-year difference-in-difference models. In section 3, I present high-level results using multiple comparison groups and specification. Finally, section 4 outlines conclusions and recommendations for policy-makers and future research.

3.1 The structure of the Mission Possible program

Mission Possible (MP) is a multifaceted program to attract and retain effective teachers and administrators in the hard-to-staff schools in Guilford County School District (GCS), North Carolina.[39] GCS is large, diverse, and predominantly low-income. In 2013-14, GCS operated 28 high schools with 23,389 high school students. The student population is 40 percent black, 12 percent Hispanic, and 6 percent Asian. Approximately 60 percent of students received free/reduced price lunch.

The MP program, which was launched in September 2006, provides recruitment, retention and performance incentives for certain teachers in a school (see further description below); staff development; performance accountability; and structural support. The county initially funded the program in 20 of its lowest-performing schools, including nine high schools. Subsequent funding from the federal Teacher Incentive Fund enabled the program to be added to another ten schools for the 2007-08 school year. To select the participating schools, the district calculated a

[39] Mission Possible Overview, http://www1.gcsnc.com/depts/mission_possible/background.htm accessed on June 1, 2014.

weighted average of each school's percent of students receiving reduced or free meals, Adequate Yearly Progress (AYP) and ABC status[40], and teacher turnover rate. As of 2013-2014, the he program was still in operation in GCS.

3.1.1 Recruitment, Retention, and Performance Bonuses

Originally, teachers and principals recruited to MP schools needed to be considered "highly qualified" by the federal requirements. In addition, they were screened and interviewed through a process modeled after the Haberman Star Teacher Interview, which focuses on how teachers work with at-risk-youth (Center for Education Compensation Reform, 2008). In later years of the program, teachers were considered for a recruitment award if they had two or more years of above average value-added data.

Table 1 reports the recruitment/retention bonuses provided to principals and teachers. The bonuses ranged from $2,500 to $10,000, with the highest bonuses provided to principals and teachers in the hardest-to-staff subject, Algebra I. These bonuses were paid to retain teachers and administrators at their positions in the years after they were recruited.

In addition to the recruitment and retention bonuses, each MP school offered performance bonuses to both teachers and administrators. To identify high-performing teachers within the system, the schools used the SAS Institute to calculate annual individual teacher value-added. Based upon these measures, if a teacher performed one standard error above the mean, he was eligible for a Level I bonus of $2,500. If instead the teacher performed 1.5 standard errors above the mean, he was eligible for a Level II bonus of $4,000. High-performing principals were

[40] The ABC's of Public Education is North Carolina's school-level accountability system created during the 1996-97 school year. Adequate Yearly Progress is an indicator whether particular schools are making progress toward federal accountability goals.

identified through school performance; if a principal's school met AYP as defined by the state accountability standards, the principal was eligible for a $5,000 performance bonus.

In Guilford County, teacher salaries are determined largely by a statewide schedule that varies by experience level, degree attained, and National Board for Professional Teaching Standards certification status, adding a small district supplement to these schedules. I was unable to obtain salary schedules for the time of interest and instead use the schedule as of the 2013-14 school year noting that it probably incorporates inflation adjustments. Teacher salaries ranged from just over $35,000 for a teacher with no prior experience and bachelor's degree to over $66,000 for a teacher with 30 or more years of experience and an advanced degree. A teacher with 15 years of experience and a master's degree earned about $48,000.[41] For this teacher, a combined recruitment/retention and performance bonus of $14,000 would represent 30 percent of his base salary.

3.1.2 Staff Development

Teachers who received bonuses were required to participate in additional staff development. Teachers in their first year of the program needed to complete two workshops entitled, *Cooperative Learning* and *Undoing Racism.* Those participating in the second year of the program were to complete workshops on *Differential Instruction* and *Teacher Expectations and Student Achievement*. The teachers had a 13-month window to complete the professional development.

[41] Guilford County Teacher Salary Schedules, http://schoolcenter.gcsnc.com/education/components/scrapbook/default.php?sectiondetailid=264921&#teacher, accessed on June 1, 2014.

3.1.3 Performance Accountability

The district also designed a mechanism to remove consistently low performers. In particular, if a teacher had negative value-added measures for more than two years or a principal failed to meet AYP for three years, they were moved to non-MP schools (Center for Education Compensation Reform, 2008).

3.1.4. Structural Support

Finally, the district provided some additional funding to the MP schools to reduce the class size of mathematics classes and purchase technologies to aid teaching (e.g., graphing calculators) (Center for Education Compensation Reform, 2008).

3.1.5 Comparison of Incentive Design with Prior Random Control Trial Research

While other studies have not found statistically significant effects of performance-based compensation, the size and design of the recruitment, retention, and performance bonuses differ in important ways from these recent randomized experiments (Springer et al., 2010; Fryer, 2011; Marsh et al., 2011; Springer et al., 2012). Similar to MP, the Project on Incentives in Teaching (POINT) in Nashville, TN (Springer et al., 2010) offered large bonuses (ranging from $5,000 - $15,000) to middle school math teachers. However, MP's incentive design differs in several important ways from POINT. First, MP focuses on the worst performing schools in the district, while POINT was specifically designed for a sample of schools with various percentiles of test score. By focusing on the worst performing schools, MP is more likely to have an effect on students' test score performance as top performing schools tend to have a score ceiling that is more difficult to shift. Second, one of the goals of MP was to place some of the best teachers in the district in front of the most struggling students. In contrast, the objective of POINT was to improve the efficiency of each teacher by incentivizing better practices. One would expect to observe smaller effects for the latter goal. Third, MP not only provided bonuses to math teachers (though they received the largest bonuses), but teachers of other subjects could qualify as well as principals. Also, many tested subject teachers were paid recruitment and retention bonuses for simply staying in the schools. As such, many teachers were guaranteed a retention bonus. Thus, MP allows for spillover effects for better leadership and non-tested subjects where POINT did not tackle these peer and leadership aspect performance.

New York City's School-wide Performance Bonus Program (SPBP) was more similar to MP (Fryer, 2011; Marsh et al., 2011). In particular, it focused on attracting better teachers to the worst performing schools in the city. However, SPBP differed from MP in two important ways. First, bonus amount were awarded at the school-level and allocation of those bonuses were decided by a committee within the school (potentially allowing for large individual bonuses for successful teachers and principals). However, often times the bonuses were distributed evenly, resulting in a smaller per person bonus of around $3,000. Second, retention bonuses were paid every year to MP school teachers, while that was not the case for SPBP teachers, possibly weakening the retention effect.

Finally, the Pilot Project on Team Incentives (PPTI) in Round Rock Independent School District (RRISD) in Texas attempted to fill the gap in the research by implementing a program that served an above average performing suburban school district with teacher team bonuses (Springer et al., 2012). While useful to better understand why other incentive programs had observed no effect, PPTI differed from the MP program in some of the ways already highlighted. First, and importantly, RRISD was already performing at the state mean. As noted with POINT, this means that potential differences between the treatment and control groups could have been smaller. Second, PPTI bonuses were smaller (less than $6,000) than MP bonuses for high school math teachers.

Taken together, the MP incentive design substantively differs from the prior randomized control trials. Thus, there is potential to find a result here despite the fact that research on other programs failed to find an effect. However, it is important to keep in mind that while the incentive design could be an improvement over these other studies, my quasi-experimental research design is weaker when drawing causal inference.

3.2. Data and methods

The North Carolina Education Research Data Center (NCERDC) collects and assembles student-level administrative data on all students in North Carolina. The data set allows students to be observed over time so that I was able to follow students from 2002 (5 years prior to the intervention) to 2010 (four years post-intervention). I linked the End of Course (e.g., Algebra I) to the End of Grade (e.g., eighth- grade math and reading) exam. I restricted the sample to ninth-grade exams. I focus on ninth grade Algebra I teachers for three reasons. First, recruiting and retaining Algebra I teachers was a particular focus of the MP program. In particular, these positions in Algebra I experienced high turnover rates and sometimes remained vacant for years. As a result, the largest teacher bonuses were awarded to Algebra I teachers. Second, the modal grade for taking Algebra I in North Carolina is ninth grade, so by examining this grade I capture when the most students take the exam. Third, very little research has focused on the effect of bonuses on students' high school math outcomes.[42] Finally, my study focuses on the high

[42] The only other study finding test score achievement effects of a recruitment bonus to low-performing schools focused on elementary schools (Glazerman et al., 2013).

schools selected during the first year of the program in order to maximize the post-intervention period.

I further restricted the sample to students who had valid eighth-grade reading and mathematics scores available[43] as well as certain demographic characteristics (e.g., age, race, free/reduced price lunch status).[44] Finally, I separated the sample into MP students and non-MP students outside of Guilford County (i.e., the rest of North Carolina). The latter category represents my comparison group.[45] Using the remainder of the state as a comparison group has the advantage of being able to find similar students throughout the state, with the downside that these students operate under different district regulations, salary structures, and other potential variations.

Table 2 shows the descriptive statistics for MP students and the comparison group in the year prior to the intervention. On average, the students in the MP schools were significantly different from the comparison group. The MP students performed worse on both the Algebra I test and the eighth-grade math and reading test. Furthermore, as expected, the MP students were more likely to be black and receive free/reduced price lunch. While the difference-in-difference model (the empirical method that I use – see description below) does not assume that the two groups are identical, merely that their trend during the pre-period are of similar slope, I estimate a basic

[43] This cut the sample size by about 10 percent. While the difference between those with and without eighth-grade math scores is statistically significant, those without the scores at the MP schools were actually higher performers on the Algebra I test. As such, any bias introduced by this sample restriction should reduce the size of the effect.

[44] Note that the administrative data contain other information about the student (e.g., English language learner), but I did not include these variables because their quality was inconsistent over time and it would have further limited the sample. It did not change the results when the model was run with more detailed demographics.

[45] Note that I also used the remaining high schools in Guilford County as a potential control group. Those effects were larger in magnitude than the findings presented here. However, given the small N, and the potential that the MP treatment had an effect on the non-MP schools (e.g., effective teachers transferring out of the school), I have not presented those findings.

propensity score to alleviate concerns about the potentially large differences between the two groups. I used the following logistic regression model:

$$\Pr(T_i = 1|X_i = x) = \frac{\exp(x'\gamma)}{1+\exp(x'\gamma)} \qquad (1)$$

where T_i is in indicator of whether the student was enrolled in an MP school (with $T_i = 1$ if student *i* was enrolled in an MP school, and $T_i = 0$ if the student was in one of the comparison schools) and X_i is a set of observable student characteristics. After estimating the propensity score, I limit the sample to an area of common support. By performing this restriction, I ensure that no comparison observation below the lowest treatment observation is included in the analysis, and no treatment observation is included above the highest comparison propensity score. Next, I weight the observations using the propensity score (for a more detailed discussion of propensity score weighting in regression, see Hirano and Imbens, 2001). Descriptive statistics on the weighted observations are shown in Table 3. The results show that propensity weights assist in making the comparison sample quite similar to the MP schools. In fact, based upon a simple t-test, there is no statistical difference between the two groups, implying that the propensity weights work well.

In order to understand the effect of MP, I first use a simple difference-in-difference model. My empirical approach is based upon the following regression:

$$Algebra_{ist} = \beta_1 Post_t + \beta_2 MP_s * Post_t + \beta_3 X_{ist} + \alpha_s + \varepsilon_{ist} \qquad (2)$$

where *Algebra* is the Algebra I test score in ninth grade standardized by year for student i, in school s, during time period t; MP_s = 1 if the student attended an MP school (the treatment group) and 0 otherwise; $Post_t$ = 1 for the years after the implementation of the program and 0 otherwise; $MP_s * Post_t$ is the interaction term yielding the effect of the policy change; X_{ist} is a vector of demographic controls including the eighth-grade reading and math test score; α_s are school fixed effects; and ε_{ist} is an error term. The interaction term between MP_s and $Post_t$ captures whether there were any changes in student test scores in the MP schools compared with the comparison group of the rest of North Carolina.

I compute the standard errors for the estimates with sandwich estimators clustering on the school. However, a number of papers have raised concerns about the appropriate standard errors for inference in difference-in-difference models when there is a small number of groups (e.g., Bertrand et al., 2004; Donald and Lang, 2007; Cameron et al., 2008; Conley and Taber, 2011). To address this issue, I adapt the method that Abadie et al. (2010) recommend for the Synthetic Control Group (and a form of the model that Bertrand et al., 2004 and Conley and Taber, 2011 recommend for difference-in-difference models). Specifically, I use exact inference with permutation tests based on placebo treatments (Abadie et al., 2010; Bertrand et al., 2004). To do this, I redo the model using a comparison district as a "placebo" treatment site. Doing this for all comparison districts gives a distribution of estimated "treatment" effects for each of the placebo treatments. If the actual estimate of the treatment effect, β_2 in model (2), is large relative to the variability in the distribution of estimated placebo treatment effects, I would conclude that the effect I observe in model (2) is larger than would be expected simply due to random variation. Put another way, the difference would be "statistically significant." Unlike conventional regression statistical inference, this procedure makes no assumptions about outcomes for students being uncorrelated over time or with other students in the same district, although it does assume that the outcomes for students are uncorrelated across districts in the same state.

A simple difference-in-difference model assumes that the trend in the pre-intervention period is similar between the treatment and comparison groups. To test this assumption, I employ a year-by-year difference-in-difference specification where I substitute lead- and lag-year dummies for the $Post_t$ indicator. This form of the interaction term will allow me to understand how the differences between the treatment and comparison groups change each year.

3.3. Results

3.3.1 Simple difference-in-difference model of Mission Possible on Algebra I test scores

The first column of Table 4 shows the basic difference-in-difference estimate of the MP program on ninth-grade Algebra I student achievement comparing the average differences between the MP schools and the rest of North Carolina. I show the classical cluster-robust standard errors in parentheses and the 90 percent “placebo confidence intervals” for North Carolina in brackets. I include school fixed effects and a set of student-level controls for eighth-grade math and reading test scores, race, age at the time of the test, various indicators for learning disabilities by subject, and a proxy for poverty.

There is some concern that the placebo confidence intervals are too conservative. In particular, the placebo confidence intervals assume that the treated schools have the same variance as the comparison schools. However, if there are significant improvements in the MP schools during the post-implementation period, this may not be a valid assumption. If the variances differ substantially, I would be less able to detect an effect. This is why I show the more classical standard errors as well as the placebo confidence intervals. In addition, I use a less conservative measure of statistical inference, a 90 percent confidence interval, when using the placebo measures of statistical significance.

The effect of the MP program compared to the comparison group is large: more than 0.10 of a standard deviation improvement over the four-year post-implementation period. However, whether or not this effect is statistically significant depends upon the measure of statistical inference. Using conventional cluster-robust standard errors, the effect is statistically significant at the 0.01 level. However, using the placebo confidence intervals, the effect falls within the 90 percent confidence interval when compared to the state. Since the placebo confidence intervals are the metric I believe to be the best, it does not appear that MP had an effect on Algebra I student achievement in the aggregate.

Recall from Table 2 that the comparison group was significantly different than the MP schools. Given the placebo method of calculating statistical inference, it could be beneficial to give greater weight to similar students, as the propensity weights do. As such, I rerun the analysis in Table 4, using the propensity weights. Generally this does not change the magnitude of the coefficients or their measures of statistical inference.

3.3.2. Year-by-year difference-in-difference specification

One disadvantage of using the change in the means in the classic difference-in-difference specification is that it could obscure an improving trend post-implementation by taking the mean over all the post-implementation years. To examine if this is the case, I employ a year-by-year difference equation disaggregating these effects. This approach also tests the assumption that no changes in test scores occurred prior to the implementation of MP.

Figure 1 shows the graphical depiction of the year-by-year regression of ninth-grade Algebra I standardized test scores on the implementation of the MP program. Specifically, each graph shows the point estimates of the MP and year interaction. The reference year is 2002. The y-axis shows the dependent variable, Algebra I test scores standardized by the annual state mean and standard deviation. The x-axis shows the spring test years, with a vertical black line through 2006, the year prior to the implementation of MP. I show both sets of confidence intervals. The traditional cluster-robust 95 percent confidence intervals are indicated by the black dotted vertical lines, while the 90 percent placebo confidence intervals are shown with a red vertical line. However, interpretation of statistical significance will differ between the two sets of confidence intervals. In the case of the traditional cluster-robust confidence intervals, I would expect to see the interval exclude zero (i.e., the estimate can be differentiated from zero) in order to be statistically significant. In the case of the placebo confidence intervals, any estimates that *exceed* the confidence intervals should be considered statistically significant. Thus, for example, if the point estimate is greater than the top of the confidence interval, then it indicates that the estimate is in the top 5 percent of all placebo estimates.

Figure 1 shows the results of the year-by-year difference-in-difference analysis using the rest of North Carolina high school students as the comparison group. These results appear to show that by the spring of 2010, MP exerted an effect on ninth-grade Algebra I test scores. In 2010, MP schools performed 0.28 standard deviations better than similar students in North Carolina. Furthermore, while the estimates are noisy during the pre-period, all the point estimates fall well within both sets of confidence intervals. While there is a slight upward trend starting in 2005, the positive trend that leads to a statistically significant result does not start until 2008. These results provide some encouragement that MP influenced Algebra I test scores.

3.3.3. Variation in the program's impacts

Although the mean impact of the program is important, part of the goal of the financial incentives and other school improvements is to improve access and subsequently performance of low-income and minority students. Even though close to three-quarters of the MP students are either low-income or black, one could imagine a scenario where more effective teachers could be tracked away from these students. In fact, Steele et al. (2014) find that even though between schools low-income minority students benefitted from more effective teachers in a couple of the sites, within a particular school that was not the case. As such, I investigate the performance of black and low-income students separately.

Figures 2 and 3 replicate Figure 1, but show the results for only black and free/reduced price lunch students, respectively. Figure 2 mimics the results for all students in Figure 1, except there does not appear to be a statistically significant effect in 2010 when I use the placebo confidence intervals (i.e., the point estimate of 0.27 is not in the top 5 percent of all observed effects throughout the state). Therefore, even though the magnitude of the effect is similar to the overall effect, I cannot conclude that black students benefited from the MP program.

Figure 3 show some encouraging signs for free/reduced price lunch students. The figure shows a similar shape as Figure 1, with no statistical difference between the MP schools and comparison group during the pre-period, and a statistically significant effect in 2010 using both sets of confidence intervals. The estimate of 0.30 is the largest magnitude of the three effects. Such results provide encouraging evidence of MP's effect on ninth-grade Algebra I achievement.

3.3.4 Evidence that the size of the recruitment/retention bonus matters

One indirect way to test if different size recruitment bonuses matter would be to see how student test scores changed on the ninth-grade English I exam compare with the results for Algebra I. Retention and performance bonuses were also provided to teachers in ninth grade English. The performance bonus was the same size as Algebra I teachers, but the recruitment/retention bonuses were much smaller (see Table 1). A problem with this approach is that the effect of recruiting/retaining lower performers cannot be disentangled from the reality that it might be more difficult to raise student achievement in English I. However, I proceed with this analysis despite this potential issue.

Figure 4 shows the performance of MP-school students on the ninth-grade English I test compared to students at other North Carolina schools by year in a similar format as Figure 1. From the figure one can see that while there was an increase in performance in 2007, the first year of the intervention, it continues to hover around zero until 2010. However, in all cases, the point estimates for the equation are well within both sets of confidence intervals indicating that we cannot differentiate the effect from zero. This figure provides some preliminary evidence that the size of the recruitment bonus matters, yet because I am using a different subject-matter test, this finding needs to be verified in future work. One piece of evidence that corroborates these findings is that the only study showing that recruiting high performing teachers to low-income elementary schools has an effect on student achievement used $20,000 bonuses (Glazerman et al., 2013). As policy-makers consider future efforts to recruit teachers and principals, the size of the bonus appears to be an important factor.

3.4. Discussion and policy implications

Teachers are one of the most important school-based factors in a child's education (see Hanushek and Rivkin, 2006), and effective administrators are an important factor in a teacher's decision to stay with a particular school (Ladd, 2011; Boyd et al., 2011; Burkhauser, 2014). MP is a comprehensive program that attempts to attract and retain high performers in schools serving high proportions of low-income and minority students in order to improve their academic achievement. The evidence presented here seems to indicate that programs like MP for high school freshmen appear to achieve the goal of improving student test scores in Algebra I after several years of implementation.

These results are encouraging especially given the size of investment for programs like MP throughout the country. The federal government through the Teacher Incentive Fund[46] as well as philanthropic dollars, most notably from the Bill and Melinda Gates Foundation,[47] have allocated more than $1 billion to encourage experimentation and incentivize states and districts across the country to measure and promote more equitable distributions of effective teachers and principals. In particular, while their study will provide a different effect than the one provided here, Mathematica Policy Research, Inc. is currently evaluating the effectiveness of MP within Guilford County. Their study randomizes 20 schools, with 10 schools participating in a performance-based compensation system while the other 10 will receive across-the-board salary raises. Further understanding the effectiveness of MP components, rather than the program as a whole, will help other districts throughout the country if they decide to create similar programs.

While encouraging, these results raise several questions for policy-makers trying to decide whether to implement a program like MP. First, can the results found in 2010 be sustained for future years? Unfortunately, the data I obtained only go through 2010, and thus I am unable to determine if the statistically significant results that I observe continue after the implementation of the bonuses. The sustainability of these effects needs to be better understood. Second, unfortunately, the data I collected from NCERDC is not detailed enough to determine a cost per pupil of the MP intervention. Being able to put teacher and principal programs like MP in the context of more recent high school research will be imperative for policy-makers when making decisions about where to invest their education dollars. For example, a different type of

[46] Teacher Incentive Fund fact sheet, http://www2.ed.gov/programs/teacherincentive/factsheet.html, accessed on June 1, 2014.

[47] Intensive Partnership for Effective Teaching press release, http://www.gatesfoundation.org/Media-Center/Press-Releases/2009/11/Foundation-Commits-$335-Million-to-Promote-Effective-Teaching-and-Raise-Student-Achievement, accessed on June 1, 2014.

intervention in Chicago (Cook et al., 2014) found larger effects of 0.65 standard deviations for a cost between $3,000 and $6,000 per pupil. Finally, while the size of the effects for these students is “large” by education standards, it is especially important to know how the program affected more long-term school consequences, like dropout and graduation rates. Performing this analysis was beyond the scope of this paper. Despite these limitations and open questions, the preliminary evidence for MP-type programs appears promising.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490).

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. *The Quarterly Journal of Economics*, *119*(1), 249-275.

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, *48*(2), 303-333.

Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2009). *Estimating principal effectiveness*. Urban Institute.

Buddin, R. (2011). Measuring Teacher Effectiveness at Improving Student Achievement in Los Angeles Elementary Schools. *Santa Monica, Calif.: RAND Corporation, unpublished working paper*.

Burkhauser, S., (2014) "Establishing the Conditions for Success: According to Teachers, Do Principals Matter in the Establishment of a Productive Working Environment?" Unpublished Working Paper.

Cook, P. J., Dodge, K., Farkas, G., Fryer R.G., Guryan, J., Ludwig, J., Mayer, S. Pollack, H. & Steinberg, L. (2014). "The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in Chicago." (No. w19862). National Bureau of Economic Research.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414-427.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education review*, *24*(4), 377-392.

Clotfelter, C., Ladd, H. F., Vigdor, J., & Wheeler, J. (2006). High-poverty schools and the distribution of teachers and principals. *NCL Rev.*, *85*, 1345.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682.

Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, *92*(5), 1352-1370.

Conley, T. G., & Taber, C. R. (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, *93*(1), 113-125.

Donald, S. G., & Lang, K. (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, *89*(2), 221-233.

Fryer, R. G. (2011). *Teacher incentives and student achievement: Evidence from New York City public schools* (No. w16850). National Bureau of Economic Research.

Glazerman, S., Protik, A., Teh, B. R., Bruch, J., & Max, J. (2013). Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. NCEE 2014-4003. *National Center for Education Evaluation and Regional Assistance*.

Greenberg, D., & McCall, J. (1974). Teacher mobility and allocation. *Journal of Human Resources*, 480-502.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of human resources*, *39*(2), 326-354.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education*, *2*, 1051-1078.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, *2*(3-4), 259-278.

Ladd, H. F. (2011). Teachers' Perceptions of Their Working Conditions How Predictive of Planned and Actual Teacher Movement?. *Educational Evaluation and Policy Analysis*, *33*(2), 235-261.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37-62.

Loeb, S., & Reininger, M. (2004). Public Policy and Teacher Labor Markets. What We Know and Why It Matters. *Education Policy Center*.

Loeb, S., Kalogrides, D., & Horng, E. L. (2010). Principal preferences and the uneven distribution of principals across schools. *Educational Evaluation and Policy Analysis*, *32*(2), 205-229.

Mansfield, R. K. (2014) "Teacher Quality and Student Inequality" Unpublished Working Paper.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., & Epstein, S. (2011). *A Big Apple for Educators: New York City's Experiment with Schoolwide Performance Bonuses: Final Evaluation Report*. Rand Corporation.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics*, *72*(2), 104-122.

Scherer, E. (2014) “Can recruitment and retention bonuses attract teachers to low performing schools? Evidence from a policy intervention in North Carolina” Unpublished working paper.

Springer, M. G., Pane, J. F., Le, V. N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team Pay for Performance Experimental Evidence From the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*, *34*(4), 367-390.

Steele, J. L., Baird, M., Engberg, J. & Hunter, G. (2014). Trends in the distribution of teacher effectiveness in the Intensive Partnership for Effective Teaching. *Unpublished manuscript*.

Steele, J. L., Murnane, R. J., & Willett, J. B. (2010). Do financial incentives help low-performing schools attract and keep academically talented teachers? Evidence from California. *Journal of Policy Analysis and Management*, *29*(3), 451-478.

Table 3.1: Incentive Structure for the Original Mission Possible Schools

	Principals	Algebra I	Mathematics Teacher (with math degree)[1]	Mathematics Teacher (without math degree)[1]	English I
Recruitment and retention bonus	$10,000	$10,000	$9,000	$2,500	$2,500
Performance bonus					
Level 1	N/A	$2,500	$2,500	$2,500	$2,500
Level 2	N/A	$4,000	$4,000	$4,000	$4,000
School meets AYP	$5,000	N/A	N/A	N/A	N/A

[1] Individuals also needed 24 hours of training in the content area that they would teach

Note: Conditions for Level 1 and Level 2 bonuses were based upon the SAS Institutes' teacher value-added. Teachers one standard error above the mean qualified for a level 1 bonus, while those 1.5 standard errors above the mean qualified for level 2 performance bonus. AYP stands for Adequate Yearly Progress as defined by the North Carolina state standards.

Table 3.2 Summary Statistics for Mission Possible (MP) and Comparison Schools in 2006 (the year prior to implementation)

Variable	MP	North Carolina Non-MP[1]	
Dependent Variable			
Algebra I	-0.94	0.01	***
Independent Variables			
8th grade math	-0.85	-0.02	***
8th grade reading	-0.76	-0.01	
% Black	72%	29%	***
% Hispanic	7%	6%	
% Asian	3%	1%	
% Other	5%	3%	
Age at time of Algebra I	15.3	15.2	***
% Free/Reduced lunch	74%	44%	***
Learning disability math	2%	2%	
Learning disability reading	7%	4%	*
Learning disability writing	6%	4%	
N	259	43,843	

[1] The "North Carolina" sample consists of all students outside of Guilford County School District (the District in which Mission Possible operates).

Note: All numbers are for 2006. Algebra I and eighth-grade math and reading scores are standardized by subtracting the state mean and dividing by the state standard deviation within grade and year. * indicates significant difference using a simple t-test at the 10 percent level, ** the 5 percent level, and *** the 1 percent level.

Table 3.3 Summary Statistics for Mission Possible (MP) and Comparison Schools in 2006 (the year prior to implementation) using propensity score weights

Variable	MP	North Carolina Non-MP[1]
Dependent Variable		
Algebra I	-0.94	-0.93
Independent Variables		
8th grade math	-0.85	-0.84
8th grade reading	-0.76	-0.74
% Black	72%	71%
% Hispanic	7%	7%
% Asian	3%	3%
% Other	5%	5%
Age at time of Algebra I	15.3	15.3
% Free/Reduced Lunch	74%	74%
Learning disability math	2%	2%
Learning disability reading	7%	6%
Learning disability writing	6%	6%
N	257	41,461

[1] The "North Carolina" sample consists of all students outside of Guilford County School District (the District in which Mission Possible operates).

Note: All numbers are for 2006. Algebra I and eighth-grade math and reading scores are standardized by subtracting the state mean and dividing by the state standard deviation within grade and year. * indicates significant difference using a simple t-test at the 10 percent level, ** the 5 percent level, and *** the 1 percent level.

Table 3.4 Estimate of the effect on Algebra I standardized scores from Mission Possible students

	No Propensity Weights	Propensity Weights
MP*Post	0.14	0.15
Standard error	-0.05	-0.05
Placebo confidence interval	[-0.50 , 0.27]	[-0.37 , 0.39]
N	266,976	266,976
Demographic Controls	Y	Y
School Fixed Effects	Y	Y

Note: Difference-in-difference estimates from a regression of ninth-grade Algebra I test scores with controls for eighth-grade test scores, race, and whether or not a student receives free/reduced price lunch (a proxy for poverty); an indicator for learning disability in math, reading, or writing; and an indicator for gender. Standard errors are cluster-robust sandwich estimator. Propensity weights calculated based upon a 2006 logistic regression of treatment status. Placebo confidence intervals calculated by redoing the difference-in-difference estimator by assigning a comparison district as a placebo "treatment" site. I redo this for all districts within the state of North Carolina. I then identify the confidence interval by examining the percentiles within the distribution. For North Carolina, I use 90 percent confidence intervals.

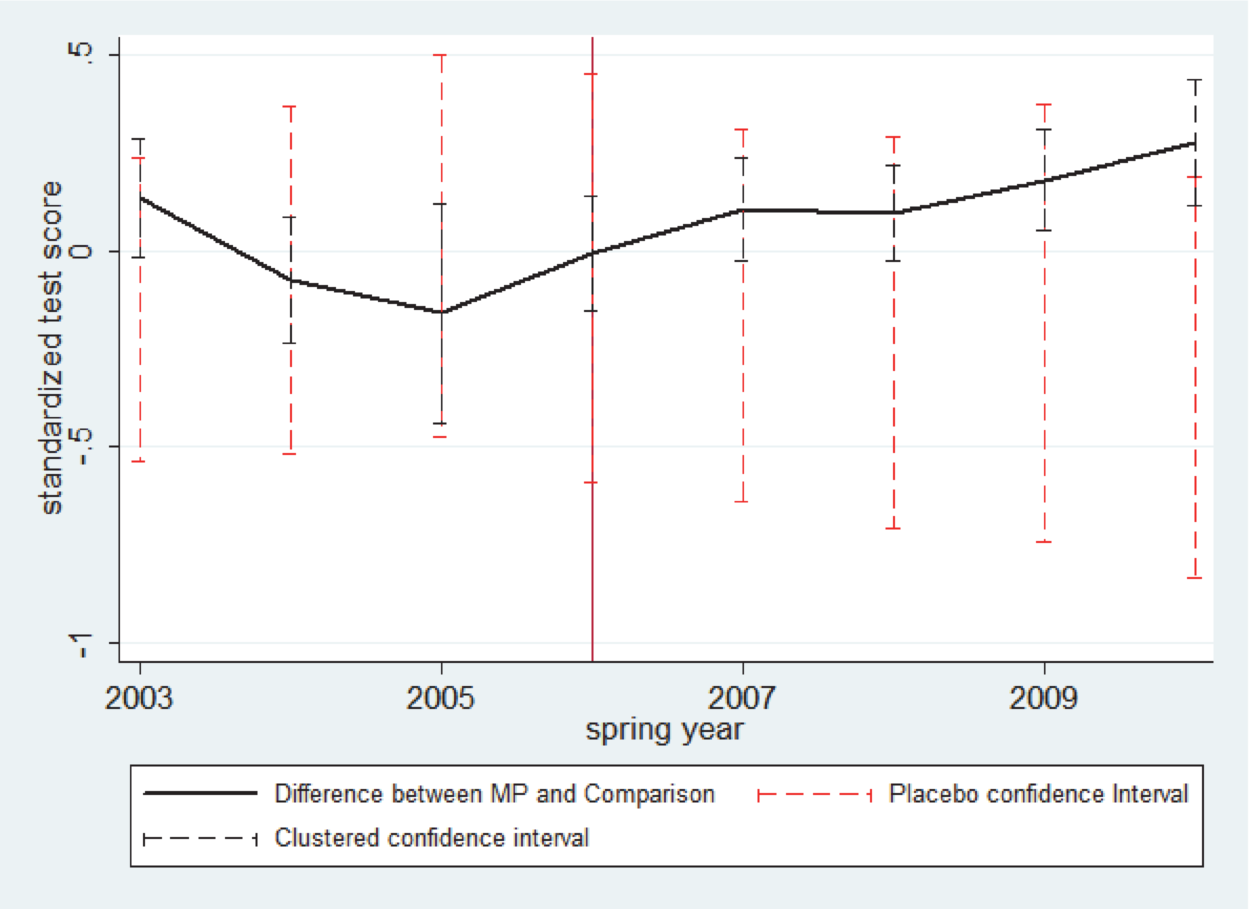

Figure 3.1: The estimated impact of Mission Possible reform on Algebra I standardized test scores. The dependent variable is the Algebra I test scores standardized by the state mean and standard deviation from 2002 through 2010. The omitted category in these estimates is 2002. Estimates are from a model that allows for effects before, during, and after the reform was adopted. The vertical black dotted lines indicate two standard errors (the 95 percent confidence intervals). The vertical red dotted lines show the 90 percent placebo confidence intervals described further in the text. All the shown estimates are weighted by the propensity score.

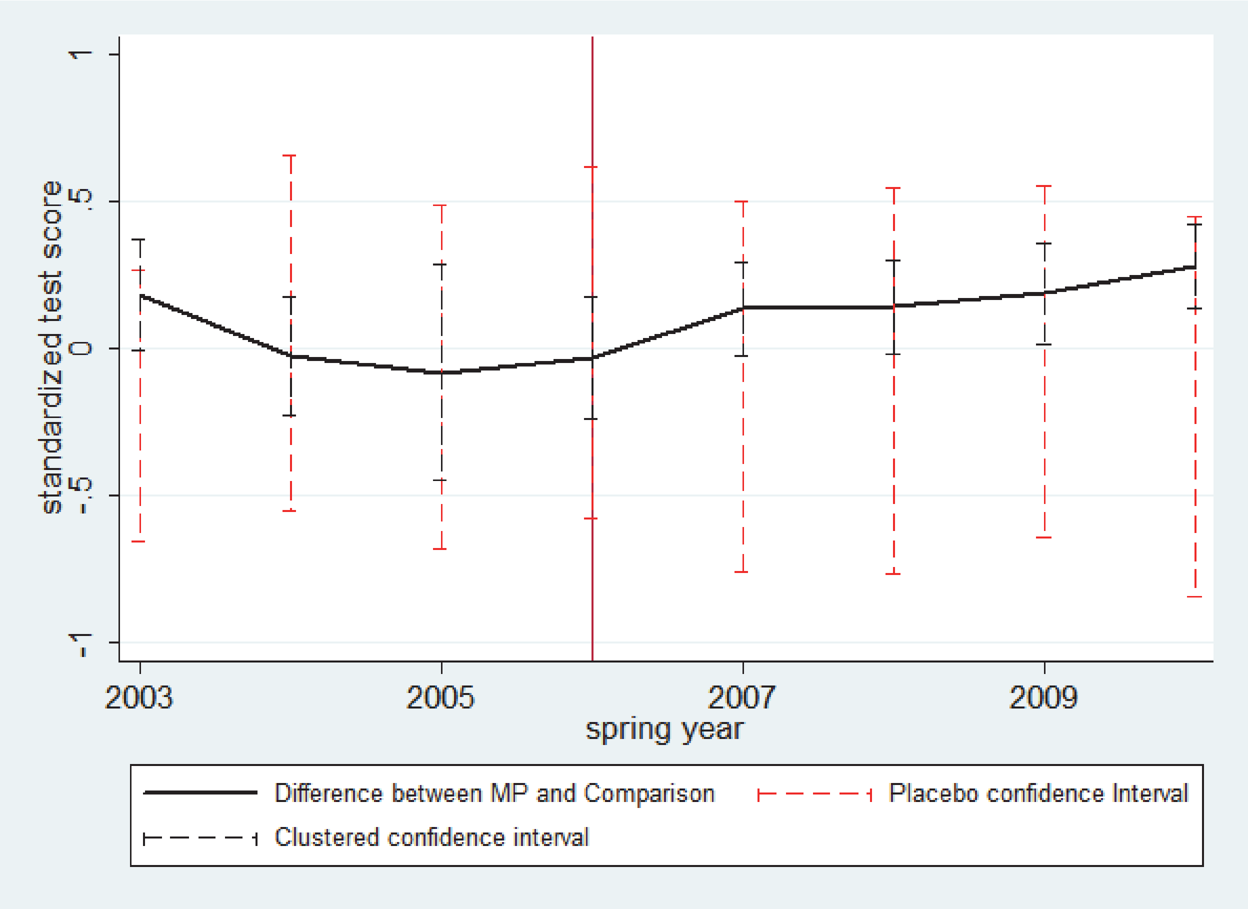

Figure 3.2: The estimated impact of Mission Possible reform on black students' Algebra I standardized test scores. The dependent variable is the Algebra I test scores standardized by the state mean and standard deviation from 2002 through 2010 for black students only. The omitted category in these estimates is 2002. Estimates are from a model that allows for effects before, during, and after the reform was adopted. The vertical black dotted lines indicate two standard errors (the 95 percent confidence intervals). The vertical red dotted lines show the 90 percent placebo confidence intervals described further in the text. All the shown estimates are weighted by the propensity score.

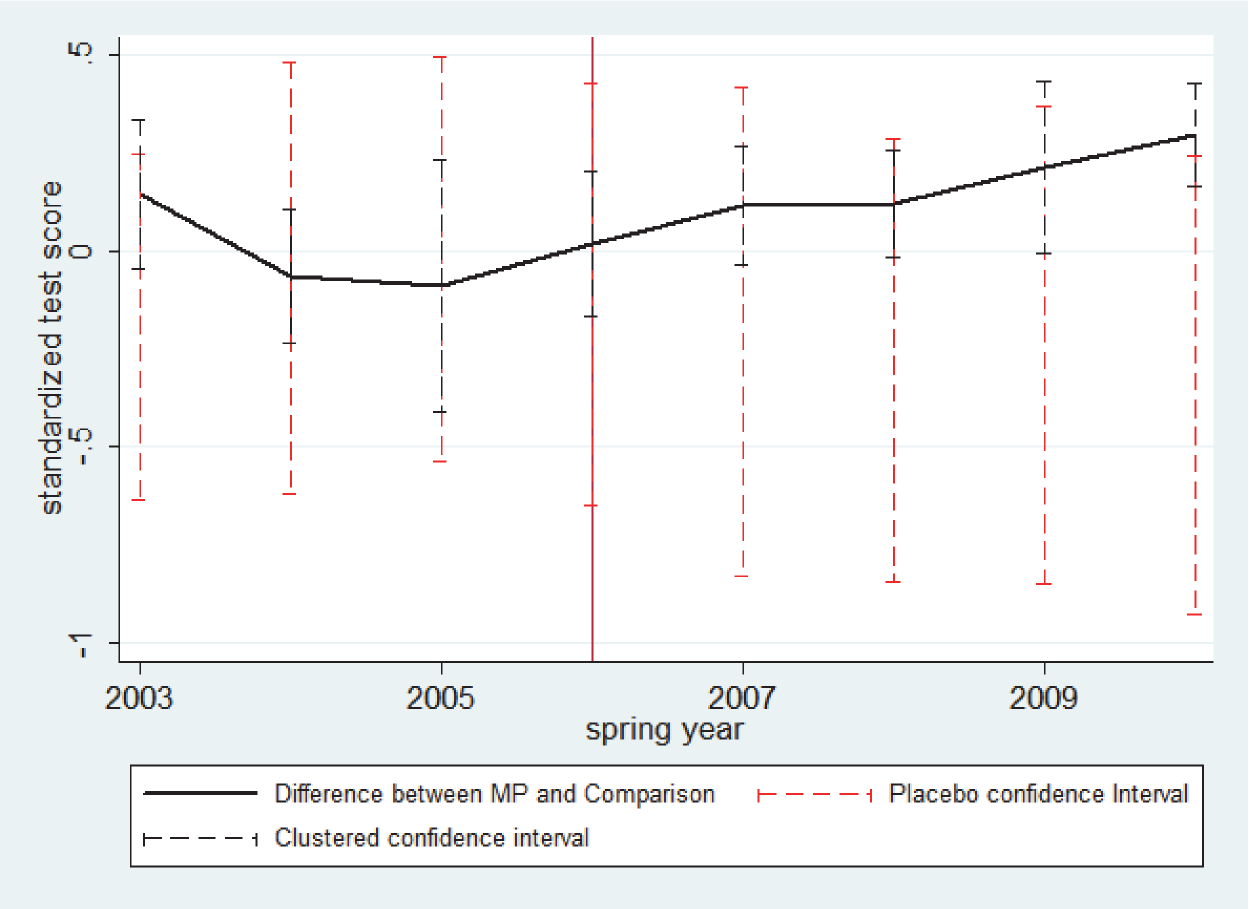

Figure 3.3: The estimated impact of Mission Possible reform on free/reduced price students' Algebra I standardized test scores. The dependent variable is the Algebra I test scores standardized by the state mean and standard deviation from 2002 through 2010 for free/reduced price lunch students only. The omitted category in these estimates is 2002. Estimates are from a model that allows for effects before, during, and after the reform was adopted. The vertical black dotted lines indicate two standard errors (the 95 percent confidence intervals). The vertical red dotted lines show the 90 percent placebo confidence intervals described further in the text. All the shown estimates are weighted by the propensity score.

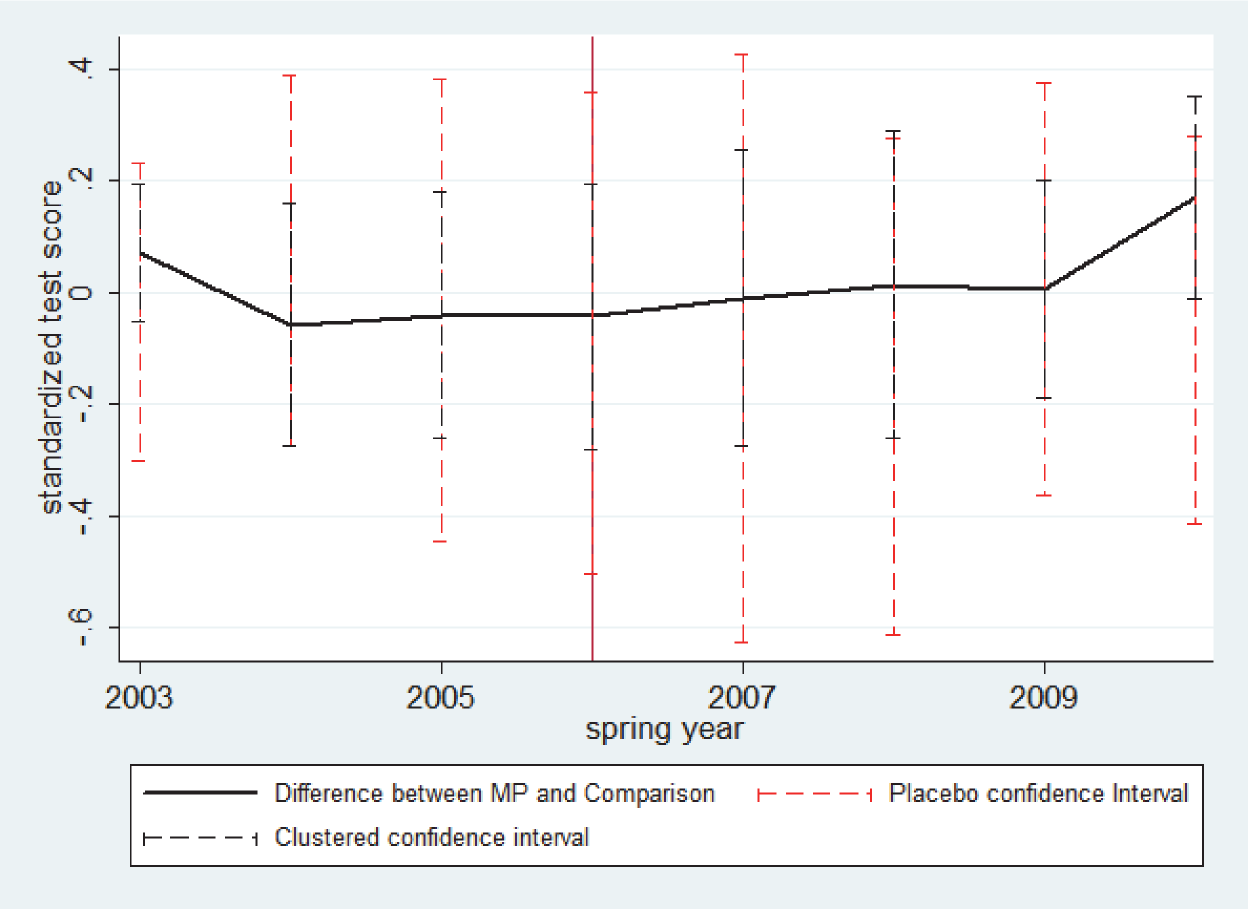

Figure 3.4: The estimated impact of Mission Possible reform on English I standardized test scores. The dependent variable is the English I test scores standardized by the state mean and standard deviation from 2002 through 2010. The omitted category in these estimates is 2002. Estimates are from a model that allows for effects before, during, and after the reform was adopted. The vertical black dotted lines indicate two standard errors (the 95 percent confidence intervals). The vertical red dotted lines show the 90 percent placebo confidence intervals described further in the text. All the shown estimates are weighted by the propensity score.

This product is part of the Pardee RAND Graduate School (PRGS) dissertation series. PRGS dissertations are produced by graduate fellows of the Pardee RAND Graduate School, the world's leading producer of Ph.D.s in policy analysis. The dissertation has been supervised; reviewed; and approved by the faculty committee composed of Jim Hosek (Chair), Paco Martorell, and Jennifer McCombs.

RAND PARDEE RAND GRADUATE SCHOOL

www.rand.org

RGSD-335