It Takes Time: The Asymmetric Impacts of Imposing and Removing a Teacher Incentive Scheme

Michael Gilraine, Uros Petronijevic, Nolan G. Pope*

July 26, 2022

Preliminary and Incomplete. Please do not cite or circulate.

Abstract

Teachers play a vital role in the education system, leading many school districts to implement payment-based incentive schemes to improve teacher quality. Despite their promise, these incentive schemes often do not generate the sought-after performance improvements. This paper sheds new light on the impact of pay-for-performance schemes by examining both the initiation *and* termination of such a program, finding a stark asymmetry between these effects. Student test scores do not change following the introduction of the program but decrease sharply (by 0.09 standard deviations) following its termination, despite the program being essentially unchanged during its ten-year tenure. The majority of the decline in test scores is explained by *within-teacher* changes in performance rather than changes to teacher composition. We argue that these differential effects arise because teachers who are responsive to incentives, preferring to work in high-stakes environments, gradually sort into program schools over time.

[*] Gilraine: Department of Economics, New York University, mike.gilraine@nyu.edu. Petronijevic: Department of Economics, York University, upetroni@yorku.ca. Pope: Department of Economics, University of Maryland, npope@umd.edu. We thank seminar audiences at Arizona State University and University of Wisconsin for helpful comments.

1 Introduction

Improving teacher quality is among the most effective ways to improve both short- and long-run student outcomes (Chetty et al., 2014). Many different polices for raising teacher quality have been proposed, including selective dismissal based on measured performance, professional development and training programs, school-based accountability programs, and pay-for-performance schemes. Payment-based incentive programs are particular appealing, given their potential to both alter the *composition* of teachers and the *effort* put forth by existing teachers (Dee and Wyckoff, 2015). Yet, despite their promise, the evidence on payment-based incentive schemes has been mixed, with many studies estimating little to no improvement in student test scores following their implementation (Atkinson et al., 2009; Springer et al., 2010).

The numerous pay-for-performance programs in existence today differ markedly across important design dimensions,[1] and while design flaws may render an incentive scheme ineffective, it may also be the case that some programs are well-designed but require time to elapse before their impact is observed. Program awareness, an understanding of its rules, and trust in the performance measures being used all take time to set in and play important roles in securing effective participation and 'buy in' among teachers (Neal, 2011). The introduction of high-stakes incentives may also alter the composition of teachers over time, with incentive-responsive or high-performing teachers sorting into such workplaces and less incentive-responsive teachers sorting out (Biasi, 2021; Leaver et al., 2021). The initial impacts of incentive programs – when these critical pieces have not yet taken hold – may therefore be quite different from the long-run impacts.

In this paper, we demonstrate the potential differences in the short- and long-run effects of teacher pay-for-performance programs by separately estimating the impact of both the

[1] Design considerations include bonus payment size (Fryer, 2013), the use group- or individual-based incentives (Goodman and Turner, 2011; Imberman and Lovenheim, 2015), the use of tournaments or fixed thresholds for evalauting performance (Loyalka et al., 2019), the reliability of performance metrics (Brehm et al., 2017), and whether there also exist punitive consequences for inadequate performance (Adnot et al., 2017).

commencement *and* termination of such a scheme. Specifically, we study the Mission Possible program, a pay-for-performance program implemented in the 2005-06 academic year and terminated in the 2016-17 year in the Guilford County Schools district in North Carolina. We couple this policy variation with detailed school administrative data, allowing us to estimate changes in teacher composition and within-teacher performance in response to both the start and end of the program.

We find a stark asymmetry between the initial effect of the program and the effect of its termination. In line with many prior studies (Glazerman and Seifullah, 2010; Springer et al., 2010, 2012; Fryer, 2013; Chiang et al., 2015; Brehm et al., 2017),[2] a difference-in-differences approach reveals little to no effect of the program on student math scores up to three years after its introduction. Ten years after its implementation, however, the program was terminated due to lack of funding. Here our assessment differs remarkably: The elimination of Mission Possible reduced student math scores by 0.09 standard deviations, with nearly the entire impact being observed immediately (i.e., in the year after) the program ended. Further, more than half of the observed decline in student math scores is driven by *within-teacher* quality reductions, with the remainder explained by changes in teacher composition in schools that were previously enrolled in the program.

The asymmetry we find occurs despite the incentive program featuring no major changes during its tenure. We also highlight that Mission Possible *immediately* offered *very* large financial incentives (up to $6,000 per year or over 10% of average annual salary) for teachers who achieved sufficiently high value-added on the job. In spite of these large incentives, the full effects of the program still took time to realize.[3] These findings underscore that even well-designed programs with substantial pay-for-performance incentives require time to reach their full potential.

Such programs are often multi-faceted and complicated, and participants require guidance

[2] United States based studies finding positive effects include Sojourner et al. (2014), Balch and Springer (2015), Dee and Wyckoff (2015), Imberman and Lovenheim (2015), and Eren (2019), while studies of successful programs abroad include Lavy (2002), Lavy (2009), and Muralidharan and Sundararaman (2011).

[3] We discuss the details of the Mission Possible program in Section 2.

on what is required of them and time to fully understand and buy into program rules. The Mission Possible program, for example, organized several information sessions over the first few years where program organizers explained and regularly reinforced the key ideas behind value-added performance measurement. These efforts gradually convinced teachers of the validity and fairness of value-added measurement and the genuine intentions of the program designers to improve student outcomes.

The passage of time also allowed higher-performing and more incentive-responsive teachers to enter program schools. While overall turnover rates did not change differentially in program schools, we show that teacher composition did gradually change. Specifically, the difference in average *prior* (i.e., measured before Mission Possible's introduction) value-added between teachers who enter and exit program program schools increased relative to the same difference in non-program schools, consistent with higher-performing teachers selecting into high-stakes incentive environments. Further, teachers who switched into Mission Possible schools also then experienced greater *improvement* in value-added upon switching than teachers who entered non-program schools.[4]

Given that Mission Possible eventually took hold with teachers who self-selected into its incentive structure, it is unsurprising that the termination of the program caused both a departure of high-performing teachers and a reduction in effort among remaining teachers. The teachers present in Mission Possible schools at the program's end were more responsive to incentives than those who where in these schools when the program started, and likely preferred working in an environment with the opportunity to earn performance bonuses.

Our findings contribute the mixed literature on payment-based incentive programs for educators by highlighting the importance of participants having full information about program details and having choice over whether to work in a high-stakes environment. Both take time to fully realize, suggesting that at least some of the estimated small or null effects of similar incentive programs found in the prior literature are under-estimates of the true

[4] Those switching into non-program schools exhibited the improvement one would expect from teachers seeking out better match quality schools (Jackson, 2013).

program effect.

The remainder of this paper is organized as follows: the next section provides a background on the Mission Possible program, while Section 3 describes our empirical strategy and data. Section 4 presents our results.

2 Institutional Background

Mission Possible was a pay-for-performance program designed to improve teacher performance and attract and retain high-quality teachers in hard-to-staff schools. Mission Possible was implemented in the Guilford County Schools district in North Carolina in the 2005-06 academic year. In its first year of operation, the program was funded with local district and philanthropic dollars and enrolled 20 low-performing schools. In the 2006-07 academic year, the district was awarded federal funding from the Teacher Incentive Fund that allowed it to add an additional 8 schools to the program, with the district receiving another Teacher Incentive Fund grant in the 2010-11 school year allowing it to enroll another 20 schools in the program. Funding from the federal grants ran out after 2015-16. While the district was able to use local funds to continue the program one more year, the 2016-17 academic year was the last year of the program. Over its course, Mission Possible enrolled 10 high schools and 40 elementary and middle schools. Given the program's focus on hard-to-staff and low-performing schools, these schools represented the lowest performing schools in the district.

The program offered two main sources of incentives. First, one-time recruitment bonuses of $5,000 were given to teachers who joined a Mission Possible school and whose value-added over the prior two academic years was above the district average. Additional yearly bonuses of $2,500 were also given to teachers in hard-to-staff positions (mostly for math and science). Second, yearly performance bonuses were given to teachers based on their value-added:[5] Teachers with a value-added score 1 standard deviation above the district average would receive $2,000 and a teacher with a value-added score 2 standard deviations above the district average would receive $6,000. Performance rewards were doubled for middle and high school math teachers. A $1,500 per teacher schoolwide performance incentive was also available, given to all teachers in a school that 'exceeds expected growth.' A secondary math teacher could therefore receive over $15,000 per year from the program (plus the initial $5,000 recruitment incentive).

Mission Possible clearly offered incredibly generous monetary incentives for working and performing well at hard-to-staff schools. However, the program did more than just set performance bonuses. Program organizers created resources and conducted information sessions to explain how value-added measurement worked and why it was a fair way to assess performance. They also designated six "teacher leaders" at each school who could act as a point of reference for their colleagues. (These teacher leaders were given $2,000 per year for their efforts.) The program further made teachers feel supported by identifying specific areas of pedagogical concern for teachers and then offering professional development programs to assist with improving these skills.

As we demonstrate below, despite the incredibly generous performance bonuses and effort to generate buy-in among teachers, program effects still took several years to surface and only become evident upon the termination of the program.

3 Empirical Strategy and Data

3.1 Empirical Strategy

Our empirical goals are twofold. First, we want to capture the effect of the introduction and termination of the Mission Possible program. Second, we aim to decompose the impact of the program at its introduction and termination into two components: (i) within-teacher

[5] Value-added was calculated using the SAS® EVAAS® model. See Vosters et al. (2018) for more details on how this model calculates value-added.

quality changes, and (ii) teacher sorting. The former component highlights the ability of teacher incentive schemes to raise teacher quality given a *fixed* set of teachers, while the latter identifies how Mission Possible raised math scores by attracting higher quality teachers. Our two empirical goals use data at different levels of aggregation: the student-level data is used directly to capture the total effect of the program, while the data is collapsed to the teacher- or school-level when we conduct the decomposition.

Our empirical strategy consists of event-study and difference-in-differences regressions that compare students and teachers in Mission Possible and non-Mission Possible schools before and after either the introduction or termination of the Mission Possible program. We have two choices to make here: (i) defining an event window, and (ii) defining the comparison group of 'non-Mission Possible' schools. For the event window, our main results use three pre and three post periods around the event (although we also explore five-year windows around the event below). We then define the 'non-Mission Possible' schools as schools in Guilford County that are not part of the Mission Possible program. (Table A.4 shows robustness to defining non-Mission Possible schools as schools in the rest of North Carolina.)

Total Effect: To estimate the introduction of the Mission Possible program we compare the test scores of students in Mission Possible schools to those in non-Mission Possible schools in Guilford County. Since the Mission Possible program introduction was staggered over three phases, we defined each phase as an event and stack our data. Specifically, we construct our sample by creating separate datasets for each of the three phases. Since naive event study estimates can be biased when there is variation in the timing of the treatment across units and treatment effect heterogeneity (Sun and Abraham, 2021), we follow Cengiz et al. (2019) to mechanically ensure that no previously treated units enter the control group. To do so, we label schools that enter the Mission Possible program during that event as Mission Possible schools while all other schools in Guilford County are labelled non-Mission Possible schools, dropping any school that will become (or that became) a Mission Possible school.

We therefore estimate the following stacked event study model:

$$y_{ispt} = \alpha + \sum_{\tau} D^{\tau}_{pt} + \sum_{\tau \neq -1} \delta_{\tau}(MP_{sp} \times D^{\tau}_{pt}) + \psi X_{ist} + \lambda_t + \gamma_p + \epsilon_{ispt}\,, \tag{1}$$

where y_{istp} is the math score of student i attending school s for Mission Possible phase p in academic year t. The variable MP_{sp} is an indicator equal to 1 if school s becomes a part of the Mission Possible program in phase p and D^{τ}_{pt} are indicators equal to 1 if year t is τ years after (or before, if negative) the phase-in year and 0 otherwise. The vector X_{ist} consists of student-level controls including lagged test scores, demographics, and assigned grade. Finally, λ_t and γ_p are time and phase fixed effects. The coefficients of interest are the δ_{τ}; they represent the difference in math scores between Mission Possible and non-Mission Possible schools τ years after (or before, if negative) the Mission Possible phase. We normalize the coefficient at t = –1 to zero and graph the δ_{τ} in event time in the figures that follow.

For table estimates, we estimate a pre-post version of equation (1):

$$y_{ispt} = \alpha + \sum_{\tau} D^{\tau}_{pt} + \beta(MP_{sp} \times Post_{pt}) + \delta X_{ist} + \lambda_t + \gamma_p + \epsilon_{ispt}\,, \tag{2}$$

where $Post_{pt}$ is an indicator equal to 1 if year t is after Mission Possible phase p and all other variables are defined in equation (1).

We estimate the effect of the elimination of the Mission Possible program in a similar way, although since Mission Possible was terminated for all schools after 2016-17 there is only one termination phase. Therefore, we can lose the phase subscript and estimate:

$$y_{ist} = \alpha + \sum_{\tau \neq 2016\text{–}17} \delta_{\tau}(MP_s \times \mathbb{1}\{t = \tau\}) + \delta X_{ist} + \lambda_t + \epsilon_{ist}\,, \tag{3}$$

where $\mathbb{1}\{t = \tau\}$ is a year indicator, and all other variables are defined in equation (1). The coefficient at $t = 2016 - 17$ is normalized to zero. A similar pre-post version of the above equation is used for the table estimates.[6]

3.2 Impact on Teachers

To investigate the impact of teachers, we start by calculating the quality of a given teacher j. We do so using value-added methodologies. Formally, we model the achievement of student i assigned to teacher j in year t as:

$$y_{ijt} = \alpha_{jt} + \beta X_{ijt} + \epsilon_{ijt} \tag{5}$$

where y_{ijt} is the student's math score, and X_{ijt} are observed characteristics of the student (demographics, past academic performance), and α_{jt} is teacher j's contribution to test scores in year t, or simply value-added. We estimate α_{jt} via fixed effects, giving us a time-varying measure of teacher quality.[7] Importantly, we allow teacher value-added to vary over time which is crucial in our setting to observe effort responses to the incentive.

Teacher Effort: We capture *within-teacher* quality changes – which we call 'effort' – by running the same event study regression as in equation (2) but include a teacher-school fixed effect. The teacher-school fixed effect ensures that we are comparing the *same* teacher at the *same* school before and after Mission Possible is introduced or terminated. Specifically, we estimate the impact of the introduction of Mission Possible on teacher effort by regressing:

$$\hat{\alpha}_{jspt} = \psi_{js} + \sum_{\tau} D^{\tau}_{pt} + \beta(MP_{sp} \times Post_{pt}) + \lambda_t + \gamma_p + \epsilon_{jspt} \,, \tag{6}$$

where $\hat{\alpha}_{jspt}$ is our VA estimate of teacher j in year t, ψ_{js} are teacher-by school fixed effects,

[6] Specifically, we regress:

$$y_{ispt} = \alpha + \beta(MP_s \times Post_t) + \delta X_{ist} + \lambda_t + \epsilon_{ispt} \,, \tag{4}$$

where $PostMP_t \equiv t \geq 2017-18$ is a post-Mission Possible program indicator, and all other variables are defined in equation (2).

[7] We do so by employing a two-step estimation procedure that first purges the covariates X_{ijt} from equation (5) and then estimating α_{jt} using fixed effects from the equation $\tilde{y}_{ijt} = \alpha_{jt} + \epsilon_{ijt}$ where $\tilde{y}_{ijt}$ are the residualized test scores resulting from the first step. Note that we do not employ empirical Bayes methodologies; this is because they are not needed in our context as we will use our VA estimates, $\hat{\alpha}_{jt}$, as a dependent variable. (In contrast, researchers using VA estimates as an *independent* variable use empirical Bayes to reduce attenuation bias.) See Koedel et al. (2015) for a detailed review of value-added methods.

and all other variables are defined in equation (2).

Similarly, we can identify the termination of Mission Possible on teacher effort by regressing:

$$\hat{\alpha}_{jst} = \psi_{js} + \beta(MP_s \times PostMP_t) + \lambda_t + \epsilon_{jst}\,, \tag{7}$$

where $PostMP_t \equiv t \geq 2017-18$ is a post-Mission Possible program indicator, and all other variables are defined in equation (6).

Teacher Sorting: To capture the impact of teacher sorting we identify the change in teacher VA in Mission Possible and non-Mission Possible schools coming from teachers entering and exiting MP schools in a given year.

Formally, let $\hat{\alpha}_j^{-\{MP\}}$ denote the VA estimate for teacher j constructed using data from years before MP was adopted.[8] Leaving out post-MP adoption data ensure that VA does not incorporate any effort response in the teacher quality calculation. Let n_{jt} denote the enrollment of teacher j's class in period t. We then take all teachers who enter school s in period t from another school s' in $t-1$[9] and find the enrollment-weighted VA, $\hat{Z}_{st}^{enter}$, of these teachers in school s:

$$\hat{Z}_{st}^{enter} = \frac{\sum_j n_{jt}\hat{\alpha}_j^{-\{MP\}}\mathbb{1}\{st \neq s', t-1\}}{\sum_j n_{jt}}\,. \tag{8}$$

Analogously, we take all teachers who exited school s in period $t-1$ and find the enrollment-weighted VA, $\hat{Z}_{sgt}^{exit}$, that these teachers would have contributed to school s in period t had they remained:

$$\hat{Z}_{st}^{exit} = \frac{\sum_j n_{j,t-1}\hat{\alpha}_j^{-\{MP\}}\mathbb{1}\{s't \neq st-1\}}{\sum_j n_{j,t-1}}\,. \tag{9}$$

[8] We also exclude any years when the teacher was in a MP school pre-MP adoption to ensure we are purging our VA estimates from any possible influences of MP schools.

[9] The set s' also includes the option of not teaching. We therefore include teachers who enter school s but did not teach in the prior year as part of our identifying variation for $\hat{Z}_{st}^{enter}$.

The change in teacher VA in school s at time t from teacher turnover, $\hat{Z}_{st}$, is then given as the (enrollment-weighted) change in VA from entering and exiting teachers: $\hat{Z}_{st} = \frac{\sum_j n_{j,t-1} \times \hat{Z}_{st}^{enter} - \sum_j n_{j,t-1} \times \hat{Z}_{st}^{exit}}{\sum_j n_{j,t} + \sum_j n_{j,t-1}}$.

Using the school-level incentive-invariant quality differences between teachers entering and exiting school s, $\hat{Z}_{st}$, we can identify the impact of Mission Possible adoption on teacher sorting by regressing:

$$\hat{Z}_{st} = \kappa_s + \sum_{\tau} D_{pt}^{\tau} + \beta(MP_{sp} \times Post_{pt}) + \lambda_t + \gamma_p + \epsilon_{st} \, , \tag{10}$$

where κ_s is a school fixed effect and all other variables are defined in equation (6). Similarly, equation (7) can be adapted to capture the impact of Mission Possible termination on teacher sorting (i.e., simply drop the 'j' subscripts).

3.3 Data

Our data consist of detailed administrative data from the North Carolina Education Research Center (NCERDC). Our student-level data include information on all public school students in the state for the 2002-03 to 2018-19 academic years. Importantly, the NCERDC data contain unique student and teachers identifiers, allowing us to match students to their teachers and to track both students and teachers over time.

The data contain test scores for each student in mathematics and English for grades two through eight from standardized tests that are administered at the end of each school year in the state.[10] Test scores are reported on a developmental scale, which is designed such that each additional test score point represents the same knowledge gain, regardless of the student's grade or baseline ability. We standardize this scale at the student level to have a mean of zero and a variance of one for each grade-year. Student-level demographics include sex, ethnicity, socioeconomic status, English learner status, disability, and gifted status.

[10] The exception is the second grade test, which is administered at the start of the school year for students in third grade. In addition, the second grade test was discontinued after 2008-09 and is not available in either 2005-06 for mathematics nor 2007-08 for English.

Summary statistics are reported in Table 1. Column (1) shows student characteristics for all students in the sample. North Carolina has a white student plurality and a substantial black minority population (27 percent), with Hispanic and Asian students making up a further twelve and three percent of the student body, respectively. Almost half of all students are socioeconomically disadvantaged, eleven percent report having a disability, and fifteen percent are gifted. Column (2) focuses on students who can be matched to teachers and thus form the sample we use to calculate teacher value-added. These students tend to be slightly lower-performing, which is caused by high-performing middle school students taking advanced math classes that have lower teacher-student match rates in our data (as multiple teachers cover these courses over the academic year).

Column (3) and (4) then display summary statistics for students attending Mission Possible schools and Guildford County schools (excluding Mission Possible schools), respectively. As Mission Possible focused on the lowest-performing schools, there are large differences across these samples with students at Mission Possible schools being far lower performing. In addition, these students are more likely to be Black (and correspondingly less likely to be white) and socioeconomically disadvantaged. Table A.1 further shows summary statistics among these schools during the three years pre and post adoption and termination of Mission Possible.

4 Results

In this section, we first document little to no effect on test scores of Mission Possible up to three years after its introduction. We then show that, more than a decade later, the termination of the program led to sharp and immediate declines in test scores measured at around 0.09 student-level standard deviations. Having established these asymmetric responses, we argue that high value-added and more responsive-to-incentives teachers gradually sorted into program schools over time. When the program was terminated, these teachers responded negatively to the removal of high-stakes bonus payments.

4.1 Aggregate Effects of Mission Possible's Introduction and Termination

Impact of the Introduction of Mission Possible: We start by reporting the effect of Mission Possible's introduction on student outcomes. Figure 1(a) displays (residualized) math test scores around the introduction of Mission Possible. Comparing the Mission Possible schools to other schools in Guilford County, the introduction of Mission Possible at event time '0' was associated with a moderate increase in test scores in the first post-introduction period. However, this increase was similar to the increase in test scores that occurred district-wide in the same period. Using other Guilford County schools as the control group, Figure A.1(a) formalizes this point by reporting the event study coefficients (from equation (1)) that represent the difference in math scores between Mission Possible and non-Mission Possible schools years after (or before) the introduction of Mission Possible. Compared to non-Mission Possible schools in Guilford County, the introduction of Mission Possible did not create any economically or statistically significant impacts on test scores. Panel A in Table 2 underscores this point, as Columns (1) to (3) report the point estimate from the pre-post event study design given by equation (2) using a three year window around the introduction of Mission Possible. We find that the introduction of Mission Possible only increased student math scores by 0.02-0.03σ and none of the estimated effects are statistically significant at conventional levels. These results increase slightly when using a five year window to 0.04-0.05σ and are marginally significant.

Impact of the Termination of Mission Possible: Now we show the impact of Mission Possible's termination. Figure 1(b) displays (residualized) math scores around the end of Mission Possible after 2016-17. A large decline in math scores in the schools that were part of Mission Possible can be seen after 2016-17, while no noticeable drops are visible for the non-Mission Possible schools in Guilford County. In addition to this raw data evidence, Figure A.1(b) shows the result for the event study coefficients from equation (1). This event shows a decrease in math test scores for Mission Possible schools of 0.05σ one year after the termination of the program. This decline in test scores more than doubled after termination to about 0.12σ. Panel B of Table 2 reports the point estimate of the impact of termination on math scores in columns (1) to (3) using a three year window around the termination of Mission Possible.. We find that the termination of Mission Possible decreased student math scores by 0.08-0.09σ. Columns (4) to (6) show that the results are similar when using a five year window.[11]

4.2 Mechanisms: Information, Teacher Effort, and Teacher Sorting

In this subsection, we discuss how Mission Possible organizers worked hard to establish trust among teachers in the validity of VA measures and to ensure an understanding of program rules. While there was no clear improvement of pre-existing teachers in Mission Possible schools when the program started, higher value-added teachers and those more responsive to incentives gradually sorted into program schools. When the program terminated, program schools were mainly comprised of teachers who selected to work in a high-stakes environment and who were therefore likely more responsive to incentives. The termination of the program consequently resulted in a large and immediate decline in performance in response to the removal of performance bonuses.

4.2.1 Information and 'Buy In'

The administrator who organized Mission Possible put substantial effort to disseminate the relevant information, develop trust, and facilitate 'buy in' from teachers. These administrators held multiple information sessions at all the mission Possible schools to help teacher understand how the program work and to help them understand the measures being used for the incentive. In addition, in each of these schools a teacher was assigned to be the Mission

[11] Since data is only available for two year post termination this increase in the event window only increases the number of years used pre-termination.

Possible team leader and would be the point person for immediate questions and concerns within the school. In addition, the administrators held regular activities to promote the program. Through these efforts the administrators tried to develop both an understanding and trust in the value-added measure being used so that teacher would actively participate in the program.

4.2.2 Effort of Pre-Existing Teachers at Mission Possible Schools at Introduction

Given the limited math scores gains from the introduction of the program, it is no surprise that our decomposition finds little within-teacher changes in quality after the commencement of Mission Possible. In particular, Figure 2(a) shows the change in teacher VA from the year after Mission Possible was introduced relative to the year prior (among teachers at that school for both years). While a large increase occurs in Mission Possible schools, a similar sized increase occurs in the rest of the district. This suggests there is little change from this source. This can be more clearly seen in Figure A.2(a) were we find no significant change to within-teacher VA after the introduction of Mission Possible when using the rest of the district as the comparison group. In columns (4) of Panel A in Table 3 we report point estimates for the within-teacher change in VA and find no evidence of an effect.

4.2.3 Sorting by Pre-Determined Teacher Quality at Introduction

While we find little evidence of and effect on within-teacher effort for pre-exiting teachers at Mission Possible schools, the program may influence higher VA teacher to enter. Figure 3(a) reports the average VA change coming from teachers entering and leaving schools around the introduction of Mission Possible. In the first year after the introduction of Mission Possible, there is a large increase in the average VA change from entering and leaving teachers. After this change in the first year the Mission Possible schools look similar to the rest of the district and North Carolina. We formally test this effect in columns (5) of Panel A in Table 3 and Figure A.3(a). Table 3 shows that while the average VA change from teachers entering and leaving Mission Possible schools was large, these effect were not statistically significant and were only limit to the teachers who moved into the school during this window (Figure 5 and Figure A.5 show turnover rates of about 15%).

4.2.4 Sorting by Responsive to Incentives

Another way in which teachers may sort into Mission Possible schools is by there own match quality with Mission Possible schools. Teachers who know that they respond well to high stake incentives may move to Mission Possible schools and then see improvements in the VA after this move. In Figure 4 we show that increase in within-teacher VA after moving to a new school was much higher for teachers that moved to a mission possible school than other schools in either the district or state. This faster improvement when a teacher switched into a Mission Possible schools than when switching to a non-Mission Possible school suggests that teachers that are more responsive to incentives are the teacher switching into Mission Possible schools. The event study of these results can be seen in Figure A.4 using the rest of the district as the comparison group and shows that the effect is statistically significant and grows to over 0.10σ over the first three years after a teacher moves to a Mission Possible school.

4.2.5 Within-Teacher Performance Declines Upon Mission Possible's Termination

After a decade of the program operating, we have teachers who chose to be there and are likely more responsive to incentives. It is therefore, not surprising that the large math score declines that we find come from a reduction in within-teacher effort. To capture these within-teacher quality changes, Figure 2(b) shows the change in teacher VA from the year prior to Mission Possible's termination (2016-17) to the following year (2017-18) among teachers at the same school for both years. The figure makes clear that there were large VA declines for the teachers at the Mission Possible schools, while VA declines at other Guilford County schools were far more limited. Figure A.2(b) show that these effects are statistically significant and grow over two post years. More formally this can be effect can be seen in Columns (4) of Panel B in Table 3. The within-teacher VA change for teachers in Mission Possible schools compared to teachers in the rest of the district after the termination of the was -0.081σ.

Panel B reports the point estimates of the decomposition in Columns (3)-(5) of Table 3. Teacher VA declined by 0.093σ in Mission Possible relative to non-Mission Possible schools after the termination of the program, with 0.081σ of this decline coming from within-teacher VA changes and 0.027 coming from changes to teacher composition. (The two components do not account for the entirety of the VA decline as the estimated compositional changes exclude some teachers, such as those new to the profession.) Within-teacher VA changes therefore accounts for nearly the entire VA decline experienced by Mission Possible schools when the program was terminated. While there is a sorting effort, it does not account for much because we must take turnover rates into account (see Figure 5 and Figure A.5), which is only around 15%, and only two post-termination years are in the data, not leaving much time for sorting.

5 Conclusion

This paper documents the asymmetric effect of the introduction and termination of a payment-based incentive scheme. We find that student test scores see little change following the introduction of the program. In contrast, we find sharp decreases in test scores following the programs termination. This asymmetry occurs despite the program being essentially unchanged during its ten-year tenure. The vast majority of the decline in test scores following the programs termination are explained by within-teacher changes in performance. These differential effects appear to arise because teachers who are responsive to incentives, preferring to work in high-stakes environments, sort into program schools over time. They then response negatively when the incentive program is ended. These findings suggest a penitential role for payment-based incentive scheme for the subset of teacher who prefer a higher stakes setting.

References

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff (2017), “Teacher turnover, teacher quality, and student achievement in dcps.” *Educational Evaluation and Policy Analysis*, 39, 54–76.

Atkinson, Adele, Simon Burgess, Bronwyn Croxson, Paul Gregg, Carol Propper, Helen Slater, and Deborah Wilson (2009), “Evaluating the impact of performance-related pay for teachers in england.” *Labour Economics*, 16, 251–261.

Balch, Ryan and Matthew G Springer (2015), “Performance pay, test scores, and student learning objectives.” *Economics of Education Review*, 44, 114–125.

Biasi, Barbara (2021), “The labor market for teachers under different pay schemes.” *American Economic Journal: Economic Policy*, 13, 63–102.

Brehm, Margaret, Scott A Imberman, and Michael F Lovenheim (2017), “Achievement effects of individual performance incentives in a teacher merit pay tournament.” *Labour Economics*, 44, 133–150.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer (2019), “The effect of minimum wages on low-wage jobs.” *Quarterly Journal of Economics*, 134, 1405–1454.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014), “Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood.” *American Economic Review*, 104, 2633–79.

Chiang, Hanley, Alison Wellington, Kristin Hallgren, Cecilia Speroni, Mariesa Herrmann, Steven Glazerman, and Jill Constantine (2015), “Evaluation of the teacher incentive fund: Implementation and impacts of pay-for-performance after two years. ncee 2015-4020.” *National Center for Education Evaluation and Regional Assistance.*

Dee, Thomas S and James Wyckoff (2015), “Incentives, selection, and teacher performance: Evidence from impact.” *Journal of Policy Analysis and Management*, 34, 267–297.

Eren, Ozkan (2019), “Teacher incentives and student achievement: Evidence from an advancement program.” *Journal of Policy Analysis and Management*, 38, 867–890.

Fryer, Roland G (2013), “Teacher incentives and student achievement: Evidence from new york city public schools.” *Journal of Labor Economics*, 31, 373–427.

Glazerman, Steven and Allison Seifullah (2010), “An evaluation of the teacher advancement program (tap) in chicago: Year two impact report.” *Mathematica Policy Research, Inc.*

Goodman, Sarena and Lesley Turner (2011), "Does whole-school performance pay improve student learning?" *Education Next*, 11, 66–72.

Imberman, Scott A and Michael F Lovenheim (2015), "Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system." *Review of Economics and Statistics*, 97, 364–386.

Jackson, C. Kirabo (2013), "Match quality, worker productivity, and worker mobility: Direct evidence from teachers." *The Review of Economics and Statistics*, 95, 1096–1116.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff (2015), "Value-added modeling: A review." *Economics of Education Review*, 47, 180–195.

Lavy, Victor (2002), "Evaluating the effect of teachers' group performance incentives on pupil achievement." *Journal of political Economy*, 110, 1286–1317.

Lavy, Victor (2009), "Performance pay and teachers' effort, productivity, and grading ethics." *American Economic Review*, 99, 1979–2011.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin (2021), "Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from rwandan primary schools." *American Economic Review*, 111, 2213—-2246.

Loyalka, Prashant, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi (2019), "Pay by design: Teacher performance pay design and the distribution of student achievement." *Journal of Labor Economics*, 37, 621—-662.

Muralidharan, Karthik and Venkatesh Sundararaman (2011), "Teacher performance pay: Experimental evidence from india." *Journal of political Economy*, 119, 39–77.

Neal, Derek (2011), "The design of performance pay in education." *Handbook of the Economics of Education*, 4, 495–550.

Sojourner, Aaron J, Elton Mykerezi, and Kristine L West (2014), "Teacher pay reform and productivity panel data evidence from adoptions of q-comp in minnesota." *Journal of Human Resources*, 49, 945–981.

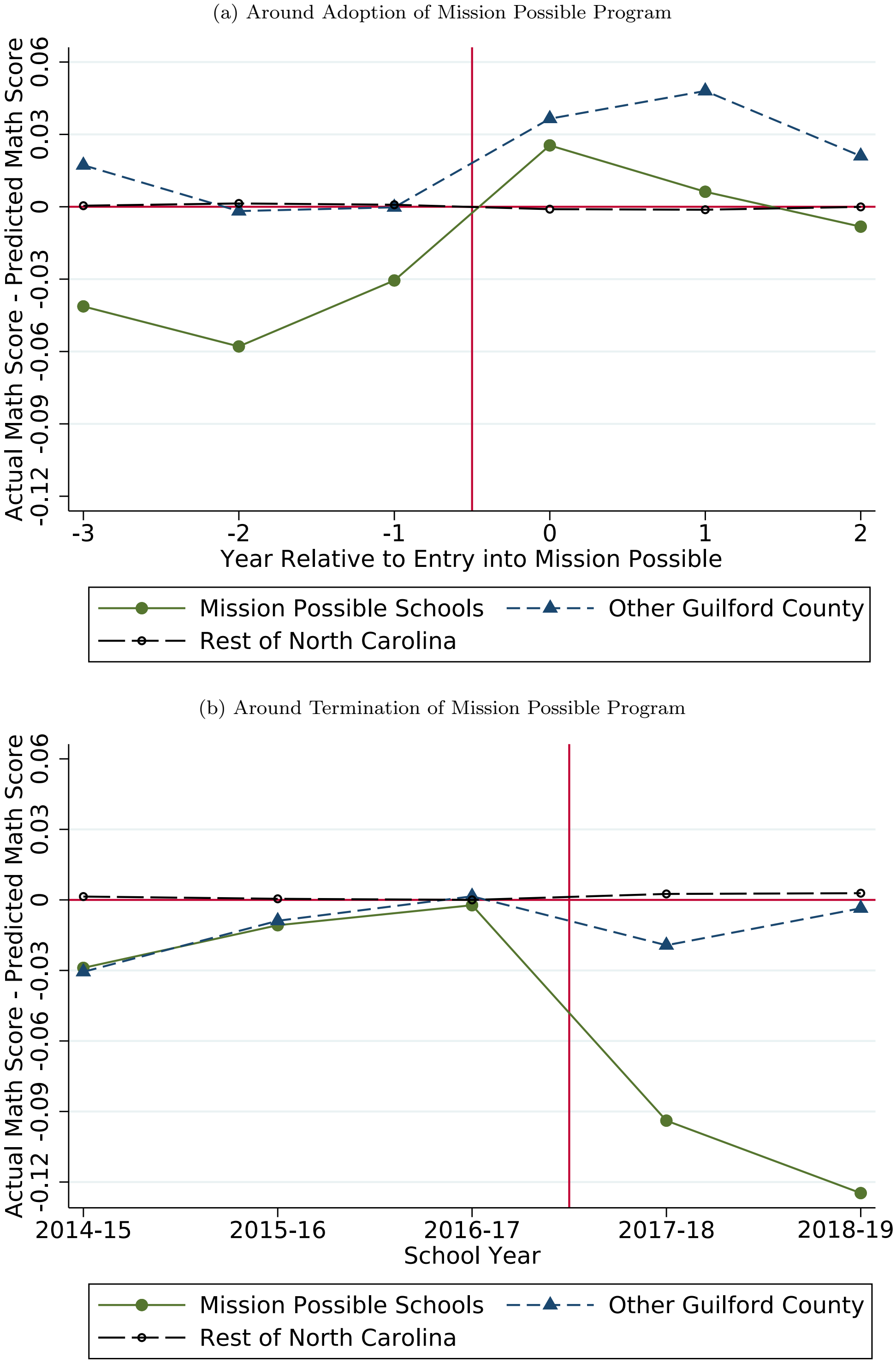
Springer, Matthew G., Dale Ballou, Laura S. Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher (2010), "Teacher pay for performance: Experimental evidence from the project on incentives in teaching." URL https://www.rand.org/pubs/reprints/RP1416.html.

Springer, Matthew G, John F Pane, Vi-Nhuan Le, Daniel F McCaffrey, Susan Freeman Burns, Laura S Hamilton, and Brian Stecher (2012), "Team pay for performance: Experimental evidence from the round rock pilot project on team incentives." *Educational Evaluation and Policy Analysis*, 34, 367–390.

Sun, Liyang and Sarah Abraham (2021), "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225, 175–199.

Vosters, Kelly N., Cassandra M. Guarino, and Jeffrey M. Wooldridge (2018), "Understanding and evaluating the sas® EVAAS® Univariate Response Model (URM) for measuring teacher effectiveness." *Economics of Education Review*, 66, 191–205.

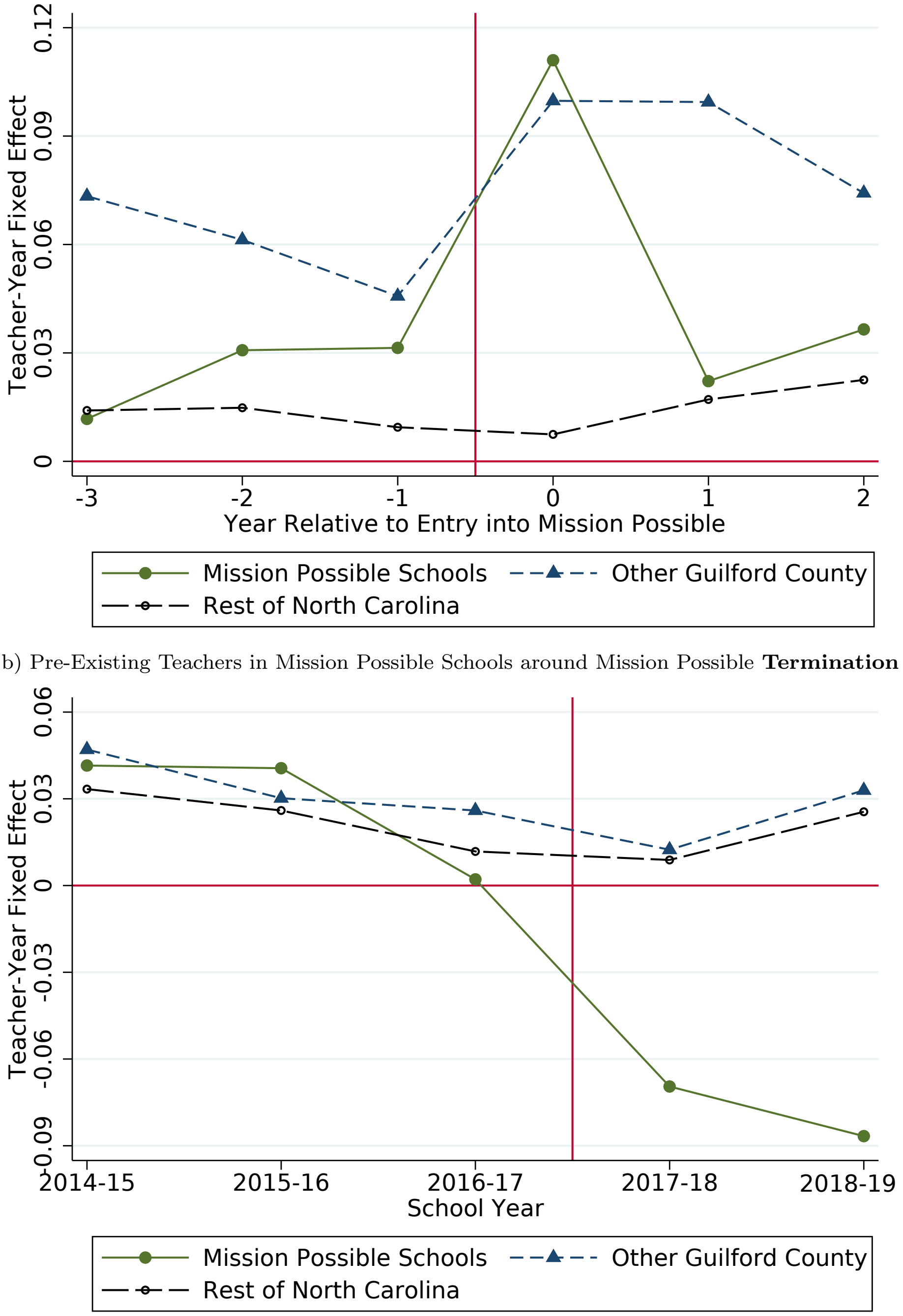
Figure 1: Impact of the Adoption and Termination of Mission Possible on Student Test Scores


(a) Around Adoption of Mission Possible Program

(b) Around Termination of Mission Possible Program

Notes: These figures show residualized test scores around the introduction of Mission Possible (Figure 1(a)) and the termination of Mission Possible (Figure 1(b)). For Mission Possible entry, we take the data from each of the three phases of program entry (in 2006-07, 2007-08, and 2010-11) and stack them and so the x-axis indicates time relative to entry into Mission Possible. Data include all 3-8 students in North Carolina with valid lagged and contemporaneous math scores. The horizontal line denotes the number zero while the vertical line indicates either when the school entered (Figure 1(a)) or exited (Figure 1(b)) the Mission Possible program. Event study versions of these figures with student fixed effects and confidence intervals are shown in Figure A.1.

Figure 2: Mechanisms: Changes in Within-Teacher Quality among Pre-Existing Teachers



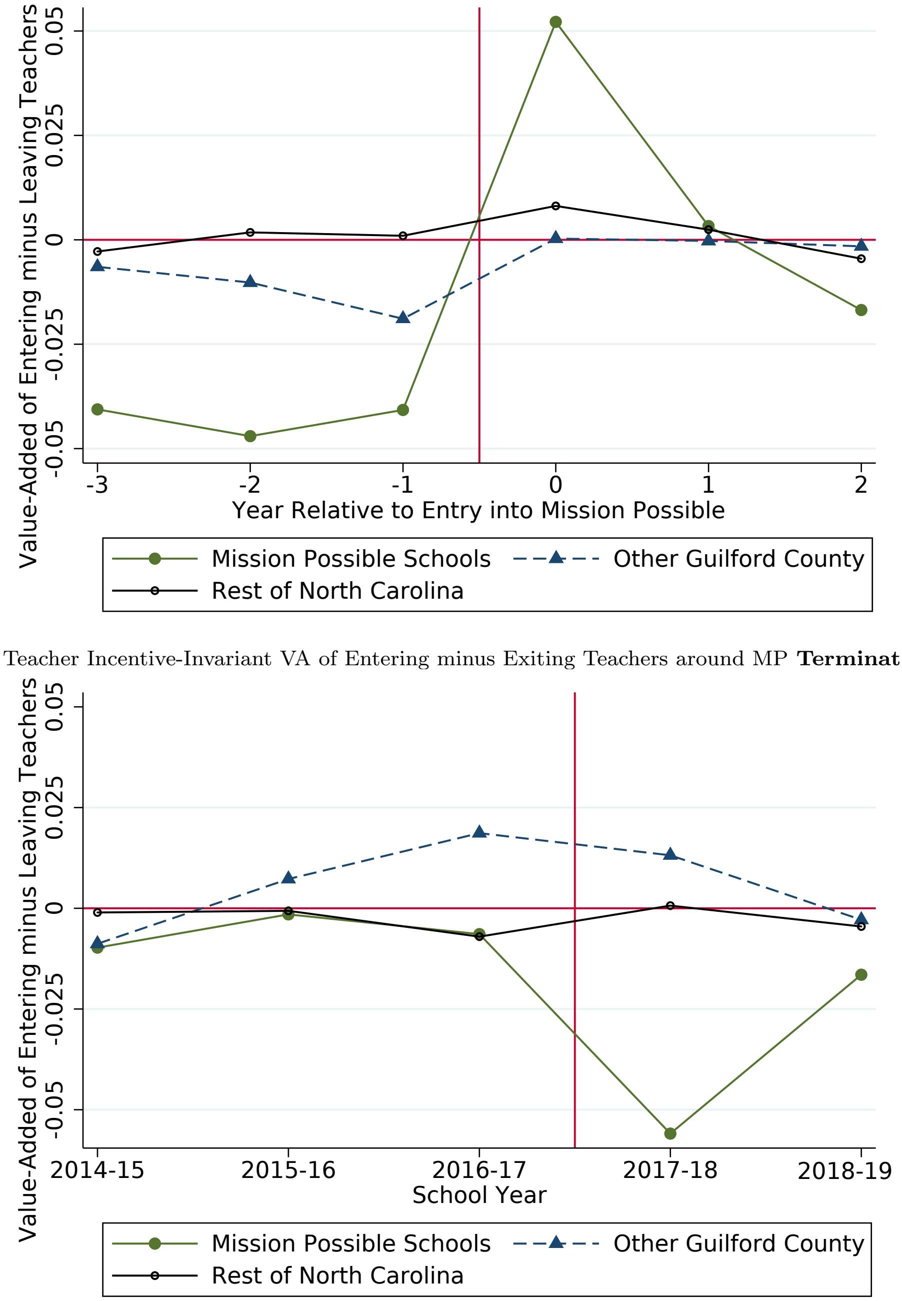
(a) Pre-Existing Teachers in Mission Possible Schools around Mission Possible **Adoption**

(b) Pre-Existing Teachers in Mission Possible Schools around Mission Possible **Termination**

Notes: These figures show mean teacher value-added for pre-existing teachers in a Mission Possible school before and after MP adoption (Figure 2(a)) or termination (Figure 2(b)). Any teacher who was at an MP school in the year prior to MP adoption/termination and was in an MP school at least once three years post-MP adoption/termination are included. For 'Other Guildford County' and 'Rest of North Carolina' pre-existing teachers are defined as those in a district (i.e., Guilford or any NC district) in the year prior to MP adoption/termination and was in a school in that district at least once three years post-MP adoption, excluding any teacher attending a MP school during the event window. For Figure 2(a) the data from each of the three phases of program entry (in 2006-07, 2007-08, and 2010-11) is stacked and so the x-axis indicates time relative to MP adoption. The horizontal line denotes the number zero while the vertical line indicates either when the school entered (Figure 1(a)) or exited (Figure 1(b)) the Mission Possible program. We note that the panels underlying these figures are unbalanced; Figure A.2, however, reports event study versions of these figures with confidence intervals and teacher-by-school fixed effects so that the same teacher is compared before and after MP adoption/termination at the same school.

Figure 3: Mechanisms: Teacher Quality of Entering and Departing Teachers **Excluding** Incentive Responses

(a) Teacher Incentive-Invariant VA of Entering minus Exiting Teachers around MP **Adoption**

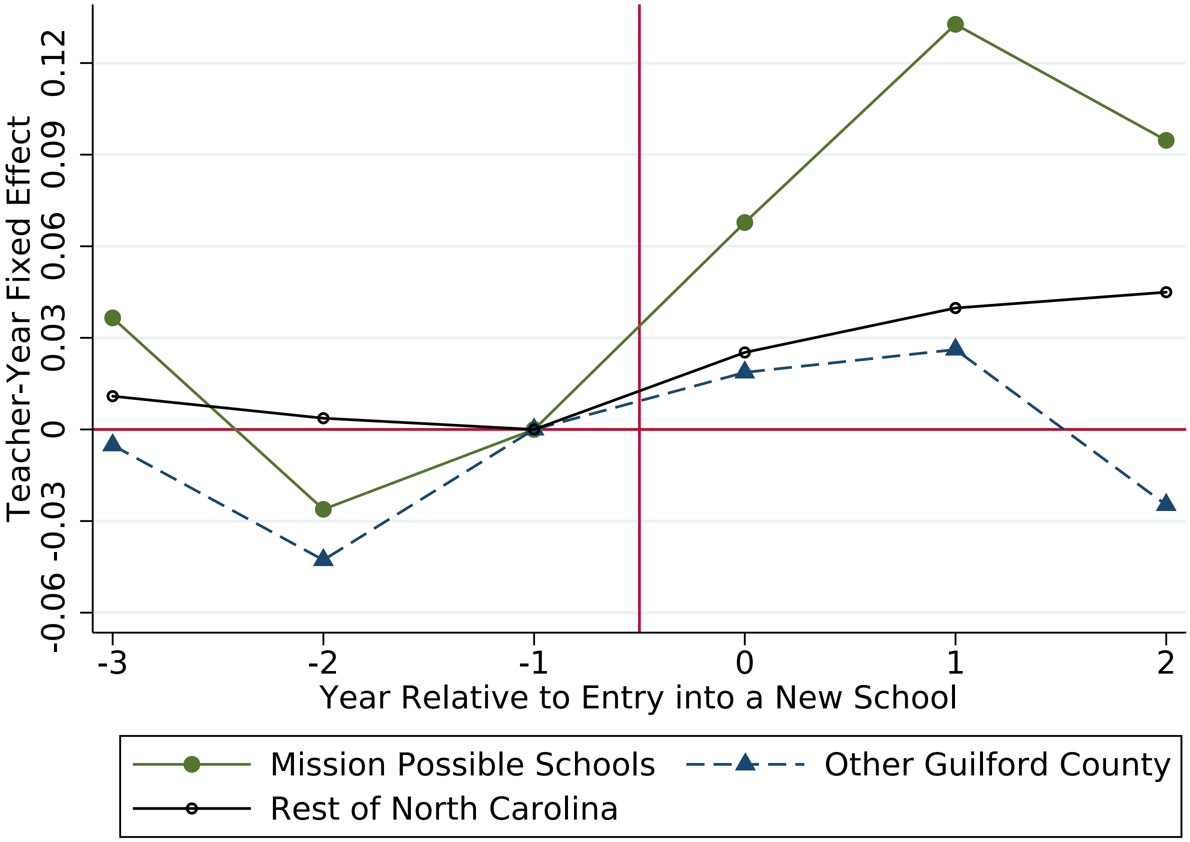


(b) Teacher Incentive-Invariant VA of Entering minus Exiting Teachers around MP **Termination**

Notes: These figures show teacher VA among teachers entering and leaving a school both around the adoption (Figure 3(a)) and termination (Figure 3(b)) of Mission Possible. Note that the VA change for a school is calculated only among those teachers entering or leaving the school and therefore does not represent the school-level teacher VA (since most teachers remain at a school in a given year). Teacher VA is calculated using only years preceding Mission Possible and also excludes any observations where a teacher taught in any Mission Possible school. These restrictions ensure that teacher VA excludes any incentive response by teachers to the Mission Possible program. The entering and exiting teacher VA is calculated using equations (8) and (9), respectively. The change in teacher VA i is then given as the (enrollment-weighted) change in VA from entering and exiting teachers. We take the data from each of the three phases of program entry (in 2006-07, 2007-08, and 2010-11) and stack them and so the x-axis indicates time relative to entry into Mission Possible. The horizontal line denotes the number zero while the vertical line indicates either when the school entered (Figure 3(a))) or exited (Figure 3(b)) the Mission Possible program. Figures are weighted by school enrollment. Event study versions of these figures with school fixed effects and confidence intervals are shown in Figure A.3.

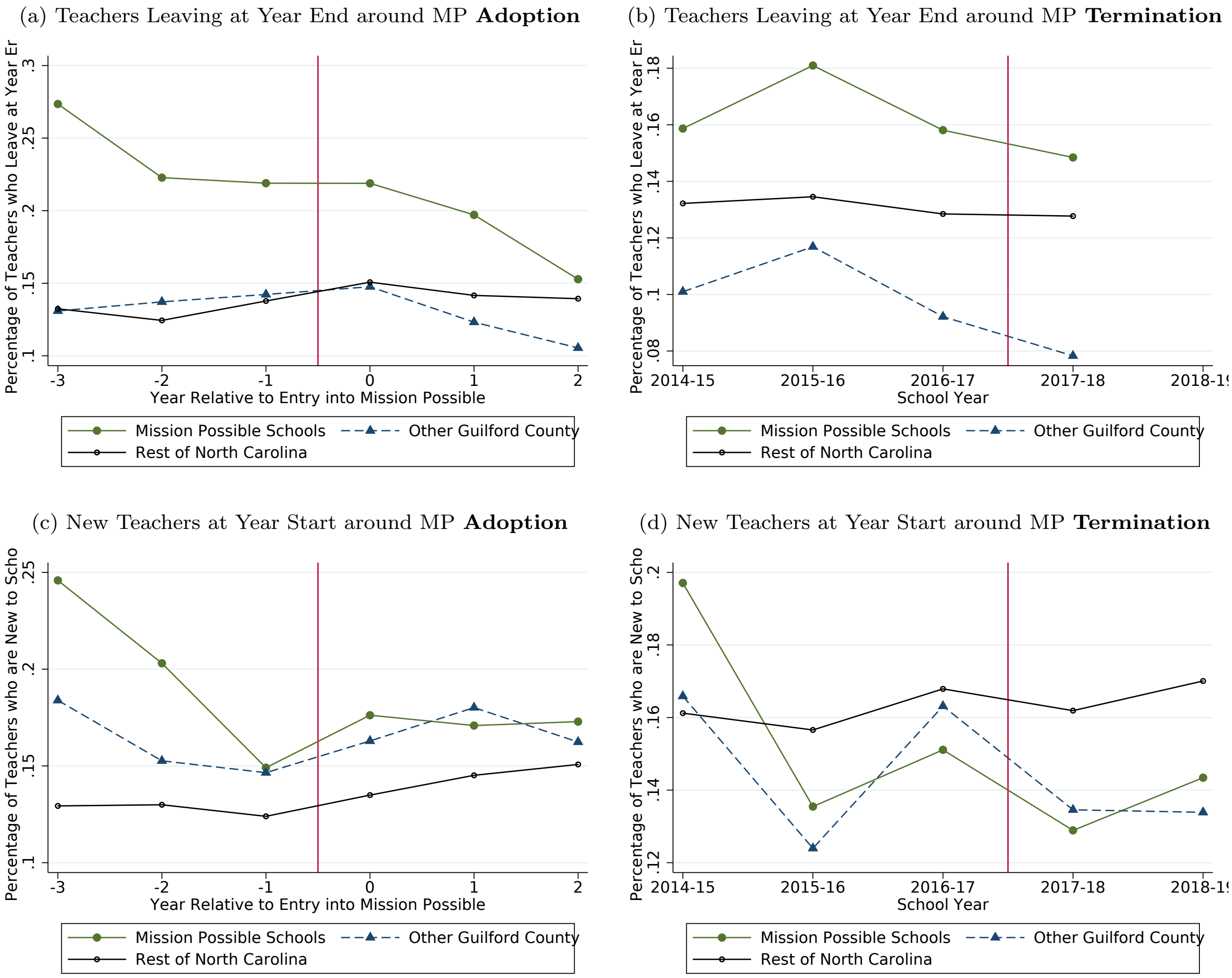
Figure 4: Mechanisms: Change in Within-Teacher Quality among Teachers that **Enter** a Mission Possible School (while Mission Possible was in operation at that school)

(a) Difference-in-Differences Point Estimate: 0.061* (0.036)



Notes: This figure conducts an event study around a teacher's entry into a Mission Possible school. The sample includes any teacher who entered a Mission Possible school while Mission Possible was in operation at that school who we observe in a non-MP school once in the prior three years. For 'Other Guildford County' and 'Rest of North Carolina' the sample covers any teacher new to the district who we observe in a different district once in the prior three years, excluding any teacher attending a MP school during the event window. Teacher VA in the year prior to entering the new school (i.e., event year '-1') is normalized to zero. The horizontal line denotes the number zero while the vertical line indicates when the teacher entered the new school. We note that the panels underlying these figures are unbalanced. We therefore also report the difference-in-differences estimate that captures the change in teacher value-added after the teacher enters a Mission Possible school, including teacher fixed effects so that the same teacher is compared before and after they enter the Mission Possible school. Teachers entering other Guilford County schools are used as the 'control' group in the difference-in-differences regression. A total of 871 teacher-year observations (covering 200 teachers) are used in the difference-in-differences regression. Standard errors clustered at school level are shown in brackets. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively. Event study versions of these figures with confidence intervals and teacher fixed effects are shown in Figure A.4.

Figure 5: Teacher Turnover Rates



(a) Teachers Leaving at Year End around MP **Adoption**

(b) Teachers Leaving at Year End around MP **Termination**

(c) New Teachers at Year Start around MP **Adoption**

(d) New Teachers at Year Start around MP **Termination**

Notes: These figures show teacher turnover rates both around the adoption (Figure 5(a) and Figure 5(c)) and termination (Figure 5(b) and Figure 5(d)) of Mission Possible. In Figure 5(a) and 5(b), the y-axis variable is calculated as the fraction of teachers in a school in the current year who are no longer present at that school the following year. In Figure 5(c) and 5(d), the y-axis variable is calculated as the fraction of teachers in a school in the current year who are new to the school – that is, who were not observed in the school the previous year. We do not observe the fraction of teachers who depart each school after the 2018-19 academic year because our data end in that year. We take the data from each of the three phases of program entry (in 2006-07, 2007-08, and 2010-11) and stack them and so the x-axis indicates time relative to entry into Mission Possible. The vertical line indicates either when the school entered (Figure 5(a) and Figure 5(c)) or exited (Figure 5(b) and Figure 5(d)) the Mission Possible program. Event study versions of these figures with school fixed effects and confidence intervals are shown in Figure A.5.

Table 1: Summary Statistics

	All of North Carolina[1]	VA Sample[2]	Mission Possible Schools	Guilford County (Excl. Mission Possible)
	(1)	(2)	(3)	(4)
Mean of Student Characteristics				
Mathematics Score (σ)	0.00	-0.02	-0.31	0.25
Reading Score (σ)	0.00	-0.03	-0.34	0.21
Lagged Mathematics Score (σ)[2]	0.01	-0.04	-0.28	0.26
Lagged Reading Score (σ)[2]	0.01	-0.03	-0.33	0.22
% White	53.4	54.5	20.5	54.1
% Black	26.9	27.0	55.0	28.4
% Hispanic	12.2	11.7	13.9	7.5
% Asian	2.6	2.1	5.8	5.1
% Economically Disadvantaged	48.7	50.5	67.3	36.0
% English Learners	5.3	4.7	8.6	4.3
% with Disability	10.5	10.4	12.7	11.3
% Gifted	15.3	13.3	16.1	31.8
# of Schools	2,385	2,261	39	53
# of Teachers[3]	59,761	59,761	1,650	1,900
# of Students	2,715,705	2,530,394	82,182	91,076
Observations (student-year)	10,045,947	7,265,433	224,438	251,008

[1] Data coverage: grades 4-8 from 2002-03 through 2018-19, grade 3 from 2002-03 to 2004-05 and 2006-07 to 2008-09. Data only includes students from cohorts that can be used to calculate teacher value-added (e.g., does not include third grade students after 2009 due to the lack of a lagged test score).
[2] Same as full sample, but restricted to grades 3-5 only for school years 2002-03 to 2005-06 due to an inability to create teacher-student matches in those grade-years. Also only includes students with non-missing current and lagged mathematics scores and in a class with more than 5 students.
[3] Only includes teachers matched to students. Also covers all teachers in those schools from 2002-03 through 2018-19 (e.g., includes teachers in Mission Possible schools before Mission Possible was implemented).

Table 2: Impact of Introduction and Termination of Mission Possible on Student Test Scores

	Event Window: 3 Years Pre and Post			*Event Window: 5 Years Pre and Post*		
Outcome: Math Scores	(1)	(2)	(3)	(4)	(5)	(6)
A. ***Introduction*** *of Mission Possible Program*						
Mission Possible Adopted	0.026	0.029	0.028	0.044**	0.049**	0.039*
	(0.023)	(0.024)	(0.021)	(0.020)	(0.021)	(0.021)
Observations	350,658	350,658	419,491	562,391	562,391	673,478
B. ***Termination*** *of Mission Possible Program (only 2 post years)*						
Mission Possible Terminated	-0.072***	-0.076***	-0.084***	-0.081***	-0.091***	-0.085***
	(0.020)	(0.020)	(0.034)	(0.021)	(0.021)	(0.033)
Controls						
Lagged Test Scores	Yes	Yes	-	Yes	Yes	-
Demographics	No	Yes	-	No	Yes	-
Student FEs	No	No	Yes	No	No	Yes
Observations	121,000	121,000	155,492	169,947	169,947	217,305

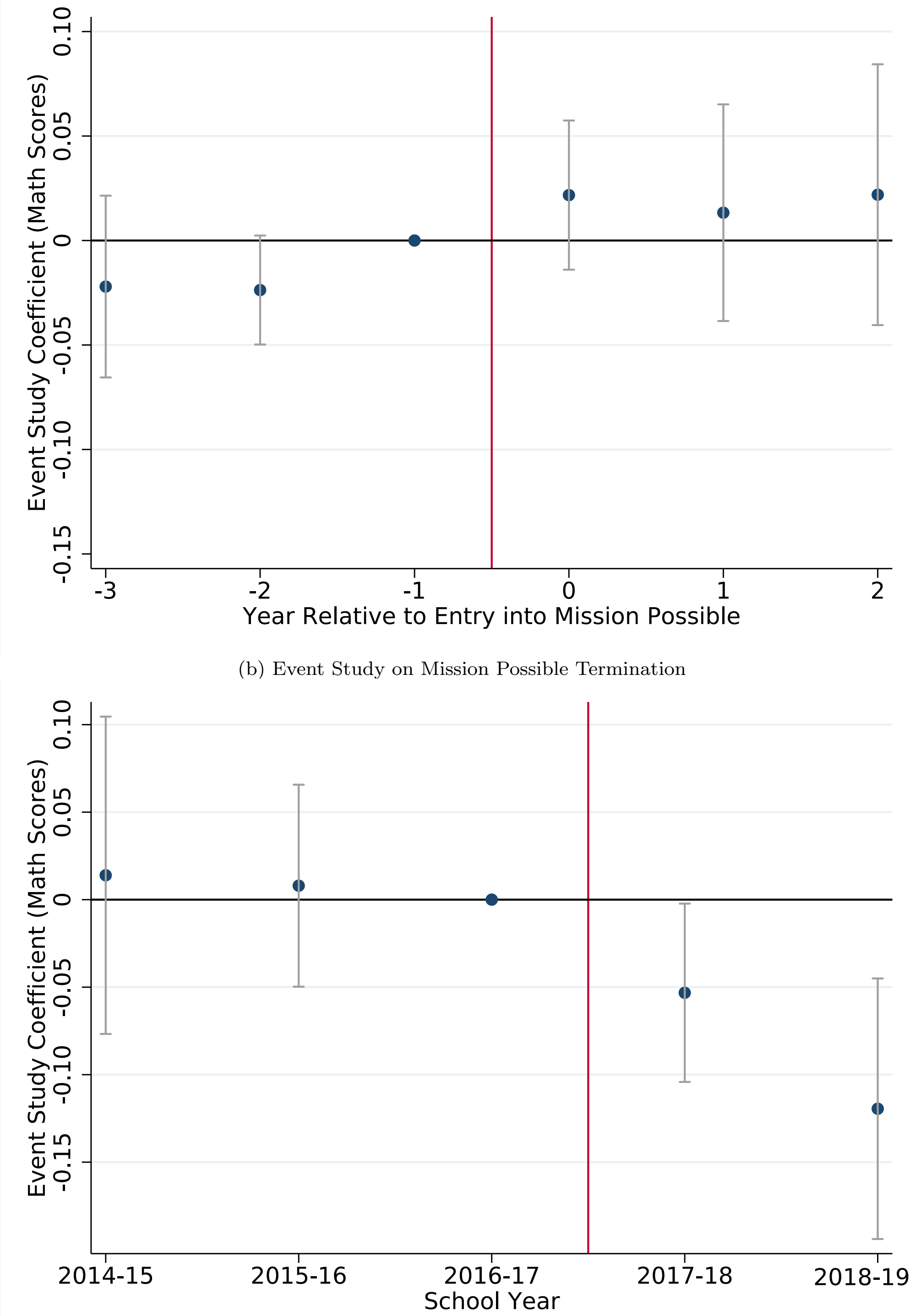
Notes: This table reports point estimates of the impact of Mission Possible on student performance when Mission Possible was adopted at a school (Panel A) and when Mission Possible was terminated (Panel B). For Panel A, we take the data from each of the three phases of program entry (in 2006-07, 2007-08, and 2010-11) and stack them as defined by equation (2). The event window around the event is defined as three years before and after adoption/termination in columns (1)-(3), while columns (4)-(6) define the event window as five years before and after adoption/termination. Note that only 2 years of post event data are available for Panel B. The outcome variable is standardized math scores. 'Lagged test scores' include cubics in lagged math and English scores, while 'demographics' include gender, ethnicity, socioeconomic status, limited English proficiency, disability status, and gifted status. Student fixed effects are used in lieu of lagged test scores or demographic controls in columns (3) and (6). All regressions include grade-by-year fixed effects. Panel A also includes entry phase and event time fixed effects. Standard errors are clustered at school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 3: Impact of Introduction and Termination of Mission Possible on Student Test Scores and Teachers

	Student-Level		*Teacher-Level*		*School-Level*
	All Students	Value-Added Sample	Teacher Value-Added	Within-Teacher VA Changes	VA Changes from Teacher Entry and Exit
	(1)	(2)	(3)	(4)	(5)
A. Introduction of Mission Possible Program					
Mission Possible Adopted	0.028	0.027	0.035	0.005	0.077
	(0.021)	(0.025)	(0.030)	(0.026)	(0.059)
Observations	419,491	192,647	7,537	3,200	787
B. Termination of Mission Possible Program					
Mission Possible Terminated	-0.084***	-0.085**	-0.093***	-0.081***	-0.027
	(0.034)	(0.044)	(0.024)	(0.030)	(0.019)
Fixed Effects	Student	Student	School	Teacher-by-School	School
Observations	155,492	103,230	2,415	1,247	475

Notes: This table show the impact of the adoption and termination of Mission Possible on students (columns (1) and (2)) and then tries to decompose this effect into the impact operating through teachers' effort (column (4)) and teacher sorting (column (5)). Column (1) simply restates the estimates from column (3) of Table 2. Column (2) then restricts the data to students we match to teachers so that it is a comparable sample to the teacher-level sample in columns (3) and (4). Column (3) then conducts the same regression as in column (2), but with the data collapsed to the teacher-level and the outcome being redefined as teacher value-added. Column (4) then investigates within-teacher changes by running a similar difference-in-differences, but using teacher-by-school fixed effects to compare teachers at the same school around the adoption of MP as described by equation (6) (Panel A) or around the termination of MP as described by equation (7) (Panel B). Column (5) then estimates compositional changes in the teaching workforce by collapsing the data down to the school-level and capturing the change in teacher value-added coming from teacher exit and entry by subtracting off the value-added of entering teachers (given by equation (8)) from the value-added of exiting teachers (given by equation (9)). Note that teachers who do not enter or exit the school are not included in the VA change calculation in column (5); since roughly 30% of teachers in MP schools attrit in a given year the impact of VA changes from teacher entry and exit on student-level outcomes is about one-third the point estimates. Standard errors are clustered at school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure A.1: Event Study: Impact of the Adoption and Termination of Mission Possible on Student Test Scores

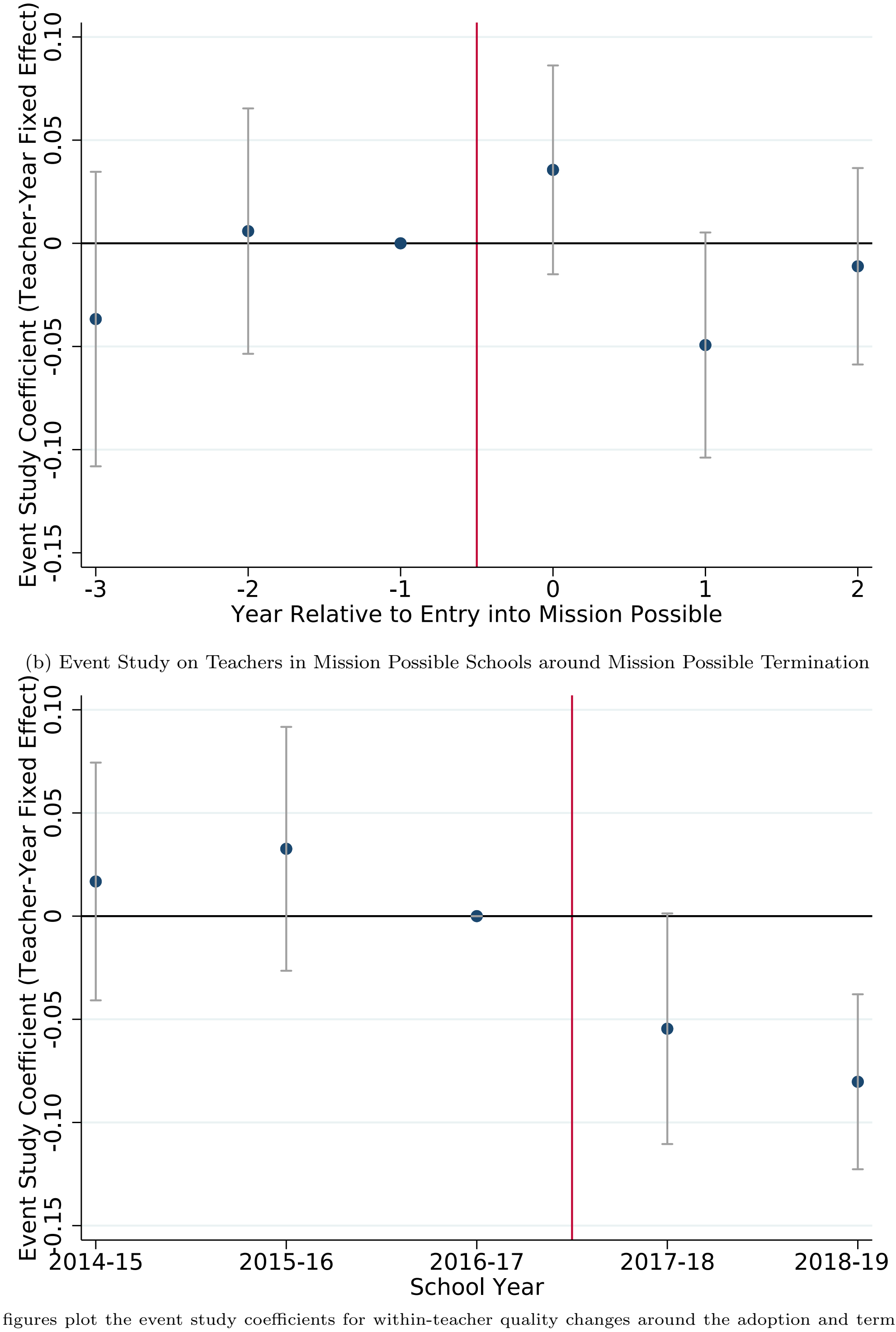


(a) Event Study on Mission Possible Adoption

(b) Event Study on Mission Possible Termination

Notes: These figures plot the event study coefficients around the adoption and termination of the Mission Possible program when other Guildford County schools are used as the control group. Effectively, they take the difference between the 'Mission Possible Schools' and 'Other Guildford County' lines in Figure 1, normalizing the point estimate at event year '-1' to zero. ('Effectively' as they also include student fixed effects.) Formally, Figure A.1(a) displays the coefficients from equation (1) around the adoption of the Mission Possible program. Figure A.1(b) then shows the coefficients from equation (3) around the termination of Mission Possible after 2016-17. The horizontal line denotes a point estimate of zero while the vertical line delineates the pre- and post-period. The whiskers represent 95 percent confidence intervals with standard errors clustered at the school level.

Figure A.2: Event Study: Change in Within-Teacher Quality



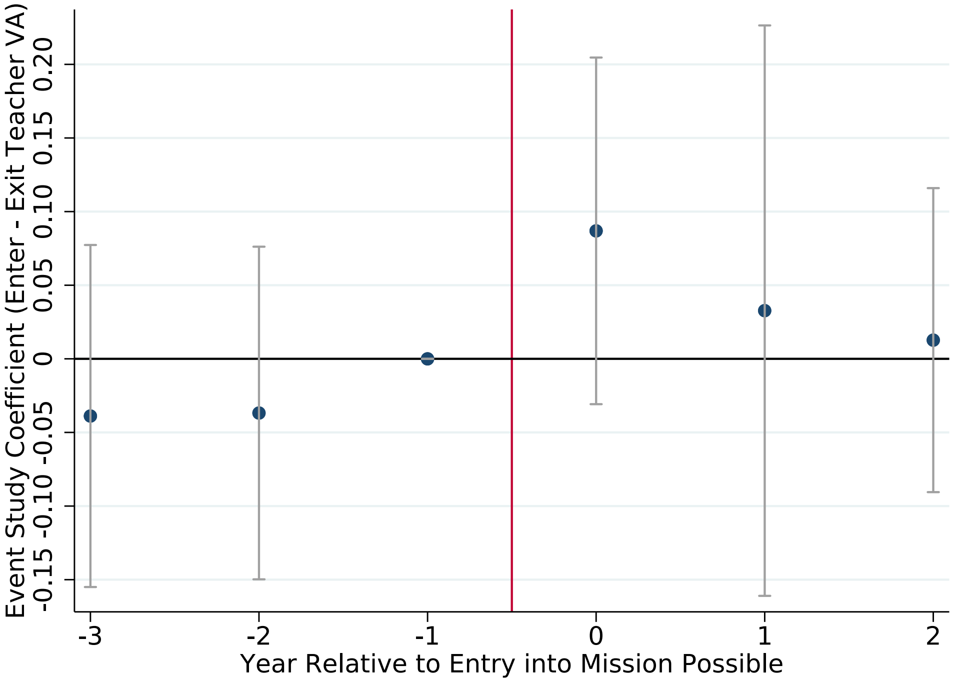
(a) Event Study on Teachers in Mission Possible Schools around Mission Possible Adoption

(b) Event Study on Teachers in Mission Possible Schools around Mission Possible Termination

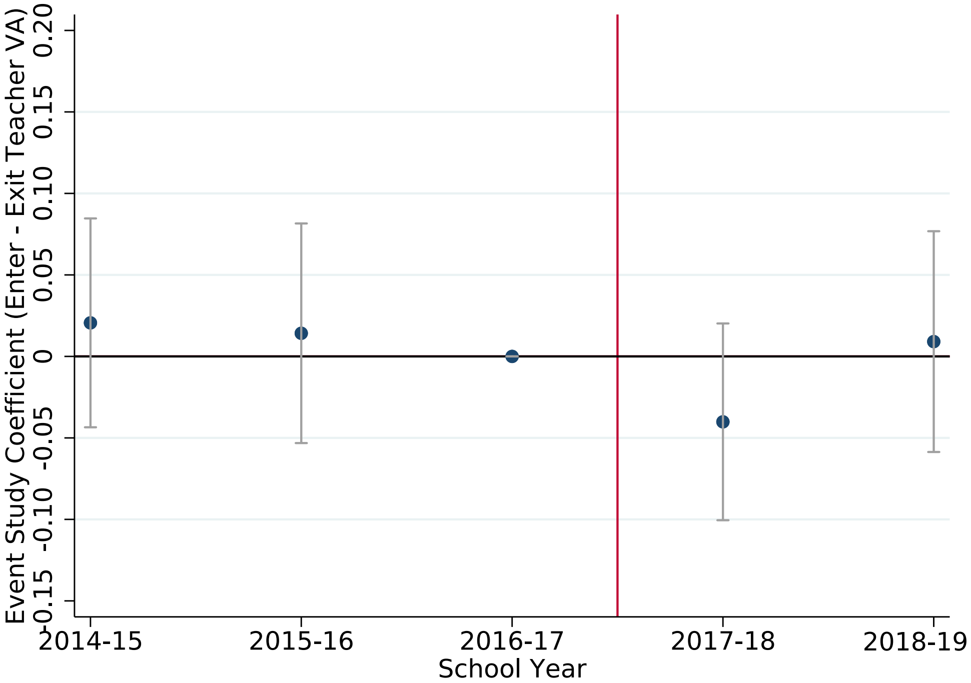
Notes: These figures plot the event study coefficients for within-teacher quality changes around the adoption and termination of the Mission Possible program when other Guildford County schools are used as the control group. Effectively, they take the difference between the 'Mission Possible Schools' and 'Other Guildford County' lines in Figure 2, normalizing the point estimate at event year '-1' to zero. ('Effectively' as they also include teacher-by-school fixed effects.) Teacher-by-school fixed effects are included so that the same teacher is compared before and after MP adoption/termination at the same school. Formally, Figure A.2(a) displays the coefficients from an event study version of equation (6) around the adoption of the Mission Possible program. Figure A.2(b) then shows the coefficients from an event study version of equation (7) around the termination of Mission Possible after 2016-17. The horizontal line denotes a point estimate of zero while the vertical line delineates the pre- and post-period. The whiskers represent 95 percent confidence intervals with standard errors clustered at the school level.

Figure A.3: Event Study: Teacher Quality of Entering and Departing Teachers **Excluding** Incentive Responses

(a) Event Study on Teacher Incentive-Invariant VA of Entering minus Exiting Teachers: MP Adoption

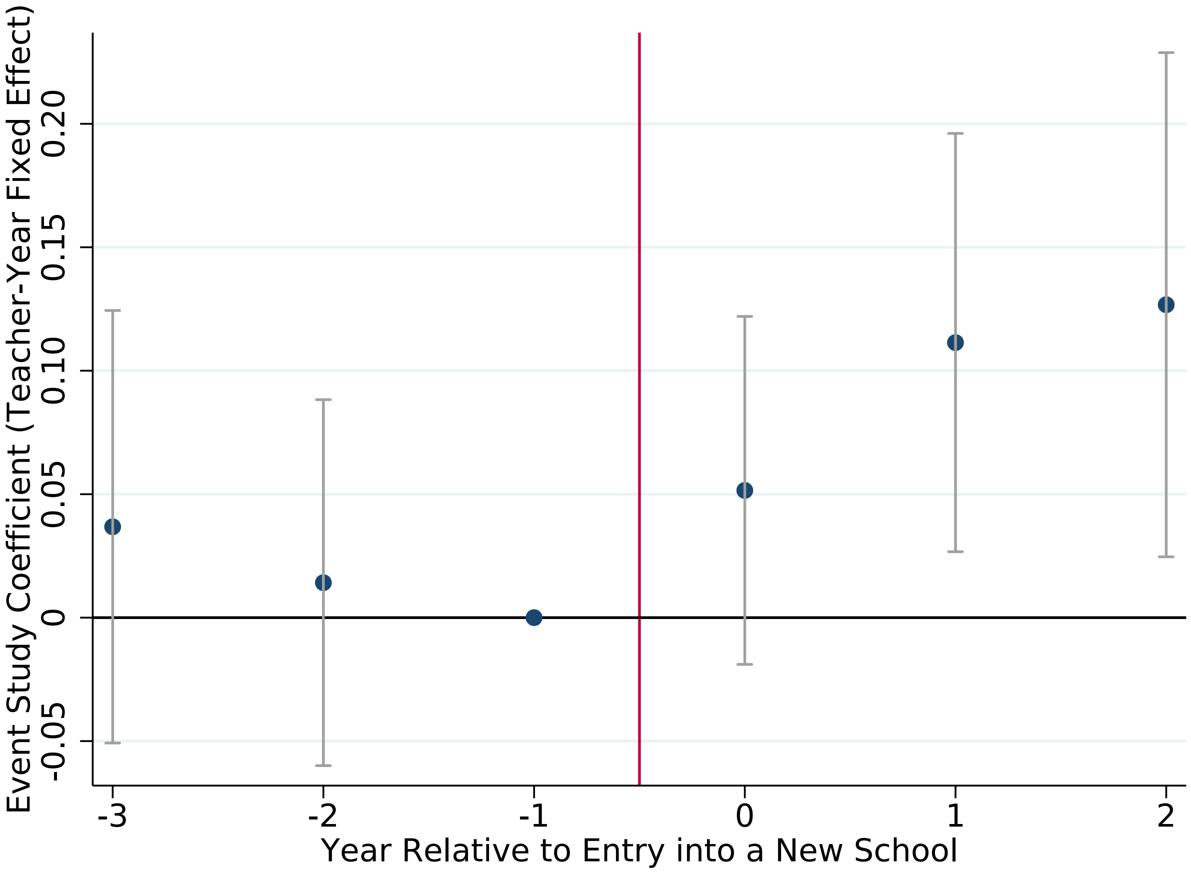


(b) Event Study on Teacher Incentive-Invariant VA of Entering minus Exiting Teachers: MP Termination



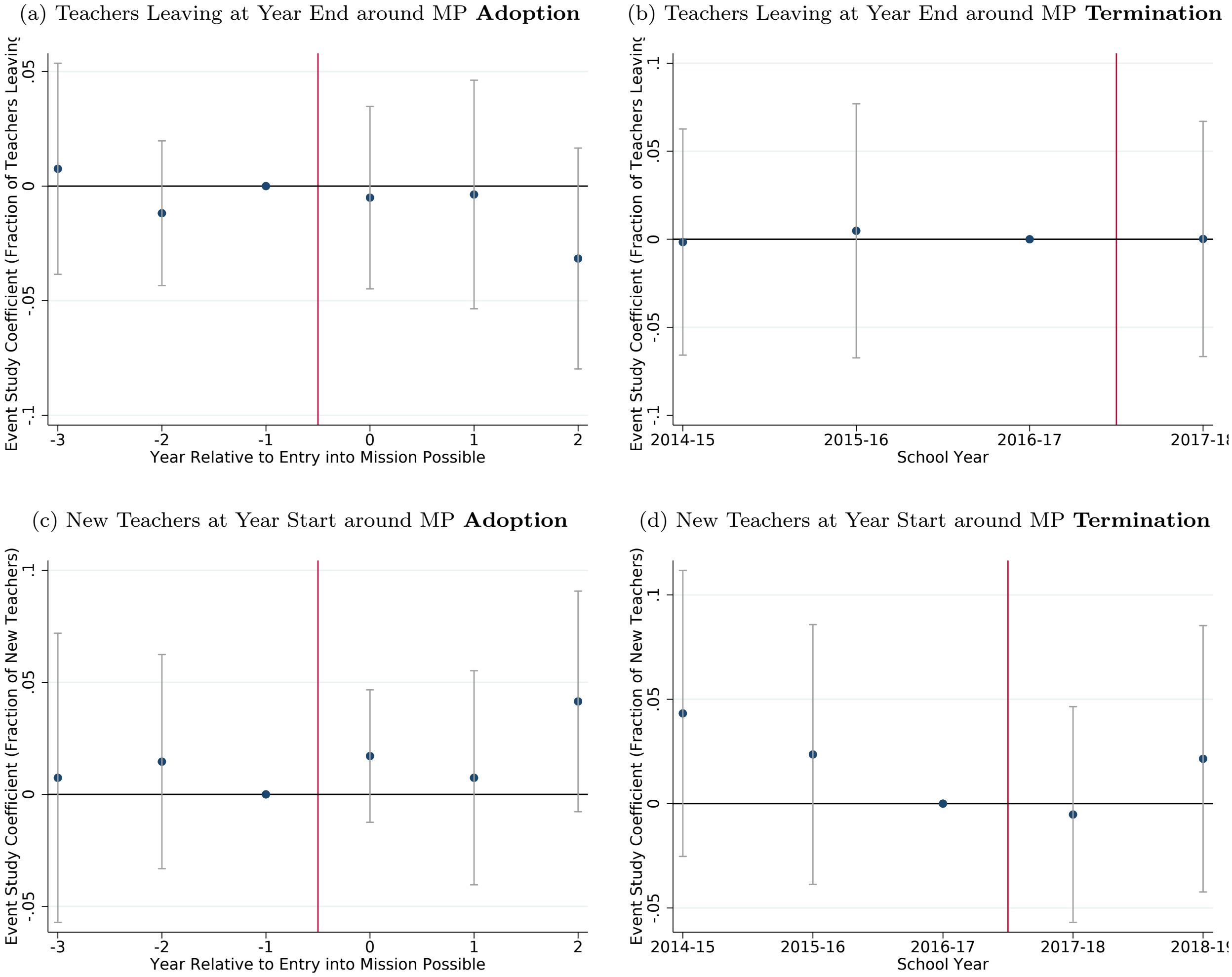
Notes: These figures plot the event study coefficients for incentive-invariant teacher quality changes for teachers who enter and depart a school. Effectively, they take the difference between the 'Mission Possible Schools' and 'Other Guildford County' lines in Figure 3, normalizing the point estimate at event year '-1' to zero. ('Effectively' as they also include school fixed effects.) Formally, Figure A.3(a) displays the coefficients from equation (10) around the adoption of the Mission Possible program. Figure A.3(b) then shows the coefficients from a similar equation (i.e., equation (7) without the 'j' subscripts) around the termination of Mission Possible after 2016-17. The horizontal line denotes a point estimate of zero while the vertical line delineates the pre- and post-period. The whiskers represent 90 percent confidence intervals with standard errors clustered at the school level.

Figure A.4: Mechanisms: Change in Within-Teacher Quality among Teachers that **Enter** a Mission Possible School (while Mission Possible was in operation at that school)



Notes: This figure plot the event study coefficients for teachers around their entry into a new school. In particular, it compares teachers that enter a MP school (while MP was in operation) to those who enter a non-MP Guilford County school from another district. Effectively, this figure takes the difference between the 'Mission Possible Schools' and 'Other Guildford County' lines in Figure 4, normalizing the point estimate at event year '-1' to zero. ('Effectively' as this figure also includes teacher fixed effects.) Teacher fixed effects are included so that the same teacher is compared before and after they enter a new school. The horizontal line denotes a point estimate of zero while the vertical line delineates the pre- and post-period. The whiskers represent 90 percent confidence intervals with standard errors clustered at the school level.

Figure A.5: Event Study: Teacher Turnover Rates


(a) Teachers Leaving at Year End around MP **Adoption**

(b) Teachers Leaving at Year End around MP **Termination**

(c) New Teachers at Year Start around MP **Adoption**

(d) New Teachers at Year Start around MP **Termination**

Notes: These figures plot the event study coefficients around the adoption and termination of the Mission Possible program when other Guildford County schools are used as the control group. Effectively, they take the difference between the 'Mission Possible Schools' and 'Other Guildford County' lines in Figure 5, normalizing the point estimate at event year '-1' to zero. ('Effectively' as they also include school fixed effects.) Formally, Figure A.5(a) and Figure A.5(b) display the coefficients from equation (1) around the adoption and termination of the Mission Possible program, respectively, when the dependent variable is the fraction of teachers who depart their school at the end of the current year. Figure A.5(c) and Figure A.5(d) display the coefficients from equation (1) around the adoption and termination of the Mission Possible program, respectively, when the dependent variable is the fraction of teachers who are new to the school in the current year. The horizontal line denotes a point estimate of zero while the vertical line delineates the pre- and post-period. The whiskers represent 95 percent confidence intervals with standard errors clustered at the school level.

Table A.1: Event Study Summary Statistics

	MP Entry Event Study Sample[1]			MP Exit Event Study Sample[2]		
	All of North Carolina	Mission Possible Schools	Guilford County (Excl. MP)	All of North Carolina	Mission Possible Schools	Guilford County (Excl. MP)
	(1)	(2)	(3)	(4)	(5)	(6)
Mean of Student Characteristics						
Mathematics Score (σ)	0.02	-0.26	0.32	0.02	-0.37	0.16
Reading Score (σ)	0.02	-0.28	0.26	0.00	-0.38	0.15
Lagged Mathematics Score (σ)[2]	0.02	-0.25	0.31	-0.02	-0.39	0.14
Lagged Reading Score (σ)[2]	0.02	-0.28	0.25	-0.02	-0.43	0.13
% White	56.4	23.5	58.5	48.8	15.3	46.9
% Black	27.8	57.0	27.2	25.4	52.3	31.5
% Hispanic	8.7	9.5	5.2	17.3	21.5	11.0
% Asian	2.1	5.3	4.3	3.1	6.3	5.5
% Economically Disadvantaged	46.9	65.3	30.7	49.2	62.6	40.5
% English Learners	4.2	6.3	2.9	4.8	8.4	3.5
% with Disability	6.8	7.8	7.3	13.1	17.2	13.9
% Gifted	15.5	16.5	32.6	15.5	16.6	33.7
# of Schools	2,225	39	53	2,127	39	56
# of Teachers[3]	43,924	1,304	1,475	24,798	608	624
# of Students	1,720,274	51,674	56188	1,060,991	27,490	31,362
Observations (student-year)	9,860,913	222,393	243,358	2,672,597	56,784	64,216

[1] Covers 3 years before and 3 years after the MP school entry for each of the three entry phases; the three entry events are then stacked. First event covers grades 3-8 from 2002-03 through 2007-08 (excluding 3rd grade in 2005-06), second event covers grades 3-8 from 2003-04 through 2008-09 (excluding 3rd grade in 2005-06), and the third event covers grades 4-8 from 2007-08 through 2012-13. Data only includes students from cohorts that can be used to calculate teacher value-added (e.g., does not include third grade students after 2009 due to the lack of a lagged test score).

[2] Covers 3 years before and 2 years after the termination of Mission Possible at the end of 2016-17. Data therefore cover grades 4-8 from 2014-15 through 2018-19.

[3] Only includes teachers matched to students.

Table A.2: Impact of Introduction and Termination of Mission Possible on Students' **English** Test Scores

	Event Window: 3 Years Pre and Post			*Event Window: 5 Years Pre and Post*		
Outcome: English Scores	(1)	(2)	(3)	(4)	(5)	(6)
A. ***Introduction*** *of Mission Possible Program*						
Mission Possible Adopted	-0.015	-0.014	0.003	-0.009	-0.004	0.005
	(0.013)	(0.013)	(0.010)	(0.012)	(0.012)	(0.009)
Observations	348,370	348,370	416,971	558,460	558,460	669,020
B. ***Termination*** *of Mission Possible Program (only 2 post years)*						
Mission Possible Terminated	-0.009	-0.010	0.018	-0.001	-0.007	0.020
	(0.012)	(0.011)	(0.016)	(0.011)	(0.010)	(0.015)
Controls						
Lagged Test Scores	Yes	Yes	-	Yes	Yes	-
Demographics	No	Yes	-	No	Yes	-
Student FEs	No	No	Yes	No	No	Yes
Observations	124,971	124,971	159,713	173,362	173,362	221,044

Notes: This table reports point estimates of the impact of Mission Possible on student performance when Mission Possible was adopted at a school (Panel A) and when Mission Possible was terminated (Panel B). For Panel A, we take the data from each of the three phases of program entry (in 2006-07, 2007-08, and 2010-11) and stack them as defined by equation (2). The event window around the event is defined as three years before and after adoption/termination in columns (1)-(3), while columns (4)-(6) define the event window as five years before and after adoption/termination. Note that only 2 years of post event data are available for Panel B. The outcome variable is standardized math scores. 'Lagged test scores' include cubics in lagged math and English scores, while 'demographics' include gender, ethnicity, socioeconomic status, limited English proficiency, disability status, and gifted status. Student fixed effects are used in lieu of lagged test scores or demographic controls in columns (3) and (6). All regressions include grade-by-year fixed effects. Panel A also includes entry phase and event time fixed effects. Standard errors are clustered at school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.3: Impact of Introduction and Termination of Mission Possible on Student Test Scores: non-Mission Possible Schools are Rest of North Carolina

	Event Window: 3 Years Pre and Post			*Event Window: 5 Years Pre and Post*		
Outcome: Math Scores	(1)	(2)	(3)	(4)	(5)	(6)
A. ***Introduction*** *of Mission Possible Program*						
Mission Possible Adopted	0.061***	0.054***	-	0.057***	0.049***	-
	(0.019)	(0.019)	(0.017)	(0.015)	(0.016)	(0.022)
Observations	9,502,462	9,502,462	9,502,462	15,174,862	562,391	562,391
B. ***Termination*** *of Mission Possible Program (only 2 post years)*						
Mission Possible Terminated	-0.103***	-0.098***	-0.093***	-0.105***	-0.106***	-0.094***
	(0.016)	(0.016)	(0.019)	(0.016)	(0.016)	(0.019)
Controls						
Lagged Test Scores	Yes	Yes	-	Yes	Yes	-
Demographics	No	Yes	-	No	Yes	-
Student FEs	No	No	Yes	No	No	Yes
Observations	2,608,381	2,608,381	3,019,920	3,632,875	3,632,875	4,289,249

Notes: This table reports point estimates of the impact of Mission Possible on student performance when Mission Possible was adopted at a school (Panel A) and when Mission Possible was terminated (Panel B) when rest of North Carolina is used as the control group (rather than Guilford County as done in Table 2). For Panel A, we take the data from each of the three phases of program entry (in 2006-07, 2007-08, and 2010-11) and stack them as defined by equation (2). The event window around the event is defined as three years before and after adoption/termination in columns (1)-(3), while columns (4)-(6) define the event window as five years before and after adoption/termination. Note that only 2 years of post event data are available for Panel B. The outcome variable is standardized math scores. 'Lagged test scores' include cubics in lagged math and English scores, while 'demographics' include gender, ethnicity, socioeconomic status, limited English proficiency, disability status, and gifted status. Student fixed effects are used in lieu of lagged test scores or demographic controls in columns (3) and (6). All regressions include grade-by-year fixed effects. Panel A also includes entry phase and event time fixed effects. Standard errors are clustered at school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.4: Impact of Introduction and Termination of Mission Possible on Student Test Scores and Teachers: non-Mission Possible Schools are Rest of North Carolina

	Student-Level		*Teacher-Level*		*School-Level*
	All Students	Value-Added Sample	Teacher Value-Added	Within-Teacher VA Changes	VA Changes from Teacher Entry and Exit
	(1)	(2)	(3)	(4)	(5)
A. Introduction of Mission Possible Program					
Mission Possible Adopted	0.051***	-	0.040*	0.048**	0.061
	(0.017)	(0.023)	(0.024)	(0.024)	(0.052)
Observations	9,502,462	5,395,516	206,257	91,389	22,673
B. Termination of Mission Possible Program					
Mission Possible Terminated	-0.093***	-0.100***	-0.081***	-0.077***	-0.029**
	(0.019)	(0.024)	(0.017)	(0.017)	(0.015)
Fixed Effects	Student	Student	School	Teacher-by-School	School
Observations	3,019,920	1,641,170	50,722	30,926	11,025

Notes: Table is identical to Table 3, but defines the control group as schools in the rest of North Carolina (i.e., not part of Guilford Country Schools) rather than non-Mission Possible schools in Guilford County. Column (1) simply restates the estimates from column (3) of Table A.3. Column (2) then restricts the data to students we match to teachers so that it is a comparable sample to the teacher-level sample in columns (3) and (4). Column (3) then conducts the same regression as in column (2), but with the data collapsed to the teacher-level and the outcome being redefined as teacher value-added. Column (4) then investigates within-teacher changes by running a similar difference-in-differences, but using teacher-by-school fixed effects to compare teachers at the same school around the adoption of MP as described by equation (6) (Panel A) or around the termination of MP as described by equation (7) (Panel B). Column (5) then estimates compositional changes in the teaching workforce by collapsing the data down to the school-level and capturing the change in teacher value-added coming from teacher exit and entry by subtracting off the value-added of entering teachers (given by equation (8)) from the value-added of exiting teachers (given by equation (9)). Note that teachers who do not enter or exit the school are not included in the VA change calculation in column (5); since roughly 30% of teachers in MP schools attrit in a given year the impact of VA changes from teacher entry and exit on student-level outcomes is about one-third the point estimates. Standard errors are clustered at school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.