The Long Run Effects of a Comprehensive Teacher Performance Pay Program on Student Outcomes*

Sarah R. Cohodes
University of Michigan and NBER

Ozkan Eren
University of California - Riverside

Orgul Ozturk†
University of South Carolina

December 22, 2023

Abstract

This paper examines the effects of a comprehensive teacher performance pay program implemented in high-need schools on students' longer-run educational, criminal justice, and economic self-sufficiency outcomes. Using linked administrative data from a Southern state, we leverage the quasi-randomness of the timing of program adoption across schools to show the school reform improved educational attainment and reduced both criminal activity and dependence on government assistance in early adulthood. We find little scope for student sorting or changes in the composition of teacher workforce to explain the findings, and instead find changes in school climate consistent with improved school environments and increased teacher efficiency. Program benefits far exceeded its costs. A comparison with similar teacher pay regimes implies that the individual incentive component of the program is necessary but not sufficient to improve student outcomes.

*We are grateful to Marianne Bitler, Katharine Parham Malhotra, Richard Mansfield, Jonah Rockoff, Eric Taylor, Yotam Shem-Tov and seminar participants at Vanderbilt University, the University of South Carolina, Clemson University and the University of Central Florida for their comments. Special thanks go to the South Carolina Revenue and Fiscal Affairs Office for facilitating the data access which made this project possible. The study was determined as exempt from human subjects review by the Institutional Review Board at the University of South Carolina.

†Cohodes: scohodes@umich.edu, Eren: ozkane@ucr.edu, Ozturk: odozturk@moore.sc.edu

1 Introduction

Improving low-performing schools is a perennial problem in education systems. Policymakers have implemented many strategies to turn around struggling schools, to varying degrees of success. One promising possibility is the use of teacher incentives. While performance pay may increase costs, it does not necessitate a large-scale hiring of staff, retraining, or a rehauling of school curriculum that may be required by other, more dramatic school reform efforts, such as a takeover by a charter management organization or state (see for example, Fryer, 2014; Abdulkadiroğlu, 2016; and Schueler, 2017). However, teacher performance pay has a mixed record in the United States, with evaluations showing negative, no, and positive impacts on test scores. This may be due to the theory of action behind incentives, the design of the incentive schemes themselves, or because prior study of such incentives has been limited to test score outcomes—measures which may not fully encompass the impacts of teachers and teacher effort (Imberman, 2015; Jackson, 2018). In analyzing the efficacy of an educational intervention, it is further important to gauge whether short-term effects (if any) persist or fade out and whether these programs spur meaningful change in long-run outcomes.

In this paper, we study the medium- and long-run effects of a comprehensive performance-based compensation program for teachers on students' educational, criminal justice, and economic self-sufficiency outcomes following the implementation of the Teacher Advancement Program (TAP) in South Carolina. TAP is a national model of teacher performance pay, which embeds incentives for teacher performance alongside professional development, the potential for career advancement, observations of teacher performance, and test-score based accountability. TAP was initially introduced in 1999 and has grown over time to serve nearly twenty states and hundreds of school districts across the U.S., the majority of which are high-need schools located in urban areas. TAP was introduced in South Carolina in 2007.

The comprehensive nature of the TAP program stands in contrast to many other teacher incentive programs which offer performance pay but little guidance on how to improve instruction to achieve thresholds for increased compensation. For example, in a randomized controlled trial of an alternative teacher incentive program in Nashville, Tennessee (POINT), teachers were offered a large, individual monetary incentive for reaching a test-score gain threshold (Springer et al., 2010). However, there were no accompanying features of the program like professional development or observations to help teachers determine *how* to increase

test scores. Instead, such a program is premised on teachers' ability to improve student performance solely by increasing effort on their own. The POINT program resulted in generally no improvement in test scores for students. Similarly, a locally-designed teacher incentive program in New York City (NYC), evaluated in Fryer (2013) and Goodman and Turner (2013), resulted in no and sometimes negative test score impacts. Again, the incentive scheme in NYC contained little guidance on how to improve student performance. In contrast to these prior evaluations of teacher incentives, there is some supporting evidence for the individual components of TAP. Consequential teacher observations and feedback can increase student test scores (Taylor and Tyler, 2012; Briole and Maurin, 2022) and test-based accountability, a form of group evaluation which may incentivize teacher performance via accountability pressure, shows gains both for student short-run outcomes (Dee and Jacob, 2011), and in some cases, longer-term outcomes (Deming et al., 2016; McElroy, 2023). However, there is little evidence that traditional teacher professional development improves student outcomes (Garet et al., 2008; Fryer, 2017; Loyalka et al., 2019) and large-scale implementation of teacher evaluation did not improve student outcomes (Bleiberg et al., 2023).[1]

In addition to its comprehensive nature, incentive pay for teachers under TAP system differs from other systems in three important ways. First, teachers' bonus allocation hinges on both their own students' achievement gains as well as the school's overall achievement growth. In this regard, TAP is a hybrid program involving both individual and group incentives, and is thereby less likely to suffer from design-specific features of one or the other pay scheme (e.g., foregone benefits in individual incentives due to lack of cooperation and free-riding in group incentives (Holmstrom, 1982; Muralidharan and Sundararaman, 2011)). Second, bonuses are substantial and sufficiently differentiated to cause changes in the behavior of educators. This is important because egalitarian distribution methods may render an incentive scheme ineffective (Fryer, 2013). Finally, teachers have the opportunity to earn bonuses based on their observed performance in the classroom *and* the resulting performance of their students. Embedding multiple measures of teacher effectiveness is a program structure choice designed to limit sub-optimal behavioral responses from teachers that may result when teacher performance is restricted to measures of student performance on standardized assessments alone (e.g., teaching to the test or neglecting the promotion of higher-order skills such as curiosity and creative thinking (Holmstrom and Milgrom, 1991)).

[1] For an overview of teacher evaluation, incentives, and training, and their connection to student outcomes, see Taylor (2023).

As part of nationwide efforts to develop and support performance-based compensation for educators in high-need schools through the U.S. Department of Education's Teacher Incentive Fund (TIF), the South Carolina Department of Education received multiple grants to implement TAP, with more than 90 schools in the state adopting the program at staggered points in time between 2007 and 2012. To identify program effects, we leverage the quasi-randomness of the timing of TAP implementation in a difference-in-differences framework using a unique data linkage from South Carolina involving administrative records from multiple state agencies spanning more than a fifteen-year period. Given the majority of TAP schools are high-need, we rely on propensity score matching to identify a set of comparison schools which are most similar to TAP schools prior to the implementation. With this strategy, in order for any observed differences in outcomes between TAP and matched comparison schools to be driven by unobservables, the timing of the change in these unobservables would have had to coincide with the timing of TAP implementation. We provide several robustness checks (e.g., tests for the existence of pretrends as well as endogenous mobility, conditioning on district-specific trends and experimenting with alternate comparison samples) and different placebo tests supporting this identifying assumption. Our main estimates employ the procedure in Borusyak et al. (2021) to address concerns about treatment effect heterogeneity biasing difference-in-difference estimates in settings with staggered treatment roll out.

We find that eighth grade students exposed to the TAP program were 3 to 4 percentage points more likely to enroll in twelfth grade and to graduate high school on time (both increases of more than 5 percent relative to the comparison means). The program also reduced students' arrest rates in adolescence and early adulthood. Specifically, students in TAP schools were 1.4 percentage points less likely to be arrested of a felony offense post-program adoption (a 30 percent decrease relative to the comparison mean). Finally, the TAP program decreased the odds of reliance on social welfare programs in (early) adulthood by 2.7 percentage points on average (a 4 percent decrease relative to the comparison mean). For all long-run outcomes, semi-dynamic treatment effects and event studies reveal a plausible dose response relationship to TAP adoption, with effect sizes growing for students exposed to TAP for a longer period in their middle school years. Being exposed to TAP is also associated with improvements in students' performance throughout their high school trajectory, as measured by both test-score and non-test outcomes. These results hold when we expand our matched comparison group to include more schools and when we use future adopters as the comparison

group.

We explore the channels through which the program was effective. The adoption of the TAP program did not change the total number of teachers in TAP schools; however, there was a small reduction in the percentage of returning teachers (roughly equivalent to one teacher per year from an average of 32 teachers). Further examination provides evidence of TAP schools attracting lower quality (measured by teacher education and a combination of state certification and teacher's subject-matter knowledge scores in state assessments) and less experienced teachers relative to school leavers. This implies that the effects of TAP reported in the paper may be lower bounds of the true impacts and that teacher sorting does not account for program effects. Indeed, we find larger program benefits when we exclude from our sample TAP schools with the highest rates of teacher turnover. We also take advantage of school climate surveys administered annually to teachers, students and parents. We find that the fraction of parents and teachers who are satisfied with learning and social and physical environments increased in the post-adoption period, although the effects for teachers are less precisely estimated. Students are not more satisfied, perhaps due to the additional effort asked of them. Taken together, our findings are consonant with explanations related to improvements in school climate as well as increases in the productivity of incumbent teachers. In an attempt to isolate the effects of components of this bundled intervention, we contrast South Carolina TAP with a version implemented in Chicago during the same time period. The test-based individual teacher incentive pay, originally integrated into the program, was omitted in Chicago TAP because of lack of reliable data for linking students and teachers. The programs were otherwise identical, yet impacts of the programs diverged. Chicago TAP did not boost student test scores (Glazerman and Seifullah, 2012), whereas, as we show, South Carolina TAP did (as well as improving the longer-term outcomes we focus on here). This comparison implies that individual incentives are a key component of the TAP intervention.

Finally, we find the TAP intervention to be cost-effective. Increases in high school graduation—despite the costs of an additional year of school to the state—alongside reductions in crime resulted in net benefits that exceeded the cost of the program. We exclude social welfare from this calculation since the costs of the program are immediate but there may be longer-term and intergenerational benefits. We calculate a marginal value of public funds (MVPF) (Hendren and Sprung-Keyser, 2020) for TAP, defined as the value of the program to recipients for every dollar spent by the government, of 14, indicating social benefit of the

program on par with that from the Abecedarian Project, a canonical preschool intervention.

This paper makes three main contributions. First, we contribute to the literature on the impacts of teacher incentives, as well as their optimal design. To our knowledge, we are the first paper in the U.S. context to investigate how teacher incentives may shape students' longer-term outcomes, rather than just test scores.[2] South Carolina TAP is a comprehensive model with multiple design features: incentives are based both on teacher practices and student performance, payments may be substantial, and the program includes feedback and support components. Evidence on TAP in other settings shows no or positive impacts on student test scores (Glazerman and Seifullah, 2012; Springer et al., 2014; Chiang et al., 2015; Eren, 2019). Non-TAP teacher incentive programs show an even wider range of impact on student outcomes, with some finding negative results (Fryer, 2013; Goodman and Turner, 2013), some positive, though typically modest, gains (Figlio and Kenny, 2007; Dee and Wyckoff, 2015; Imberman and Lovenheim, 2015; Speroni et al., 2020; Biasi, 2021; Hanushek et al., 2023; Morgan et al., 2023), and others null impacts (Sojourner et al., 2014).[3] Our findings suggest that a comprehensive pay scheme, embedded with observations of teaching practices and a feedback mechanism, can deliver desired student outcomes in a cost-effective way.

Second, we add to the evidence on school turnaround strategies more broadly. TAP was targeted to low-performing schools, which are a frequent subject of education reform efforts designed to increase student performance. Such efforts include comprehensive school reform (CSR) (Borman et al., 2003; Borman et al., 2007), which entails adoption of a school-wide curriculum and retraining teachers to implement it; adopting charter school practices (Fryer, 2014; Abdulkadiroğlu et al., 2016), which includes takeovers by charter management organizations as well as the adoption of specific practices, and state and federal school turnaround efforts, which may involve hiring new staff, state takeover of district management, extended learning time, and other initiatives (Schueler et al., 2017; Zimmer et al., 2017; Bonilla and Dee, 2020; Schueler et al., 2022). While many of these strategies result in improved student performance, these efforts all entail major revamping of school staff and practices and general upheaval within the school community,

[2]Lavy (2020) examines the effects of a performance-based compensation program for teachers, which was conducted in 49 Israeli high schools, on long-term human capital and labor market outcomes in adulthood. Schools were randomly assigned to either a treatment or a control group such that teachers at treatment schools were eligible to earn individual performance bonuses on the basis of their own students' achievement. This study shows that students exposed to treatment experienced sizeable gains in postsecondary education and annual earnings.

[3]The evidence on the impact of incentive pay on student achievement from other countries is more encouraging. See, for example Atkinson et al. (2009) for England; Glewwe et al. (2010) for Kenya; and, Muralidharan and Sundararaman (2011) for India.

all of which may make them unpalatable as large-scale reform efforts. TAP stands in contrast as a program targeted to improving school performance, but one that works with existing school staff and practices, focusing on teachers to improve student performance.

Finally, we contribute to understanding of the relationship between educational and social interventions on short-run (typically test score) and longer-run outcomes (educational attainment, criminal justice, and social welfare). A mounting body of evidence suggests that short-run effects of educational interventions can differ substantively from longer-run effects (see Bailey et al., 2017 and Bailey et al., 2020 for overviews and discussions of this phenomena). For example, researchers have shown that short-run effects do not fully capture long-run effects when examining Head Start and other preschool programs (Ludwig and Miller, 2007; Gray-Lobe et al., 2021; Anders et al., 2023), class size (Chetty et al., 2011; Dynarski et al., 2013), school choice (Deming et al., 2014; Beuermann and Jackson, 2022), accelerated learning (Cohodes, 2020), and Medicaid access for children (Cohodes et al., 2016). Our findings that modest test score gains precede meaningful increases in educational attainment, decreases in criminal activity, and reliance on social welfare programs are consistent with this pattern, and more broadly point to the importance of examining longer-run outcomes when evaluating interventions for young people.

The paper proceeds as follows. We describe TAP and how it was deployed in South Carolina in Section 2. Section 3 describes the data and empirical methodology. We follow this with Section 4, which reports results and the findings from several robustness checks. Sections 5 and 6 include a discussion of mechanisms and a benefit-cost analysis of the program, respectively. We conclude in Section 7.

2 Background

This section describes the Teacher Advancement Program and its specific implementation in South Carolina — the context for this study.

2.1 The Teacher Advancement Program

The Teacher Advancement Program (TAP) is a comprehensive school reform model designed to develop, support and retain high-quality teachers and, ultimately, improve student achievement. Since its inception in 1999, TAP has grown steadily and become one of the nation's largest education programs, serving nearly

twenty states and hundreds of school districts, the majority of which are high-need schools located in urban areas.

There are four key, interrelated elements of TAP: (i) multiple career paths, (ii) ongoing applied professional growth, (iii) instruction-focused accountability, and (iv) performance-based compensation. Multiple career paths enable skilled teachers to assume greater leadership roles without having to leave the classroom by serving as master and mentor teachers. Additional responsibilities include, but are not limited to, coaching and mentoring classroom teachers, developing research-based instructional strategies, and supporting principals in outlining the school's focus for improvement.

The second element of TAP, ongoing applied professional growth, allows teachers to learn new instructional strategies, collaborate with master and mentor teachers, and receive individual coaching. Teachers meet in grade-alike or subject-alike groups under the guidance of master and mentor teachers for about 50 to 90 minutes each week. Instruction-focused accountability, the third program component, requires teachers in TAP schools to be held accountable for high-quality instruction. Teachers are evaluated four to six times during the school year by school administrators and master and mentor teachers in different areas of effective instructional practice for an overall classroom observation score. Post-evaluation sessions are also held by observers to help teachers strengthen their instructional practices. Finally, teachers in TAP schools are eligible for additional compensation based on their performance in the classroom (teaching practices) and their students' and overall school performance (teaching outcomes).

2.2 South Carolina Teacher Advancement Program

The U.S. Congress established the Teacher Incentive Fund (TIF) in 2006 to support performance-based compensation systems for educators in high-need schools. The TIF program made five-year grants available to local and state education agencies and delivered multiple rounds of grants which included TIF 1 in 2007, TIF 2 in 2008, TIF 3 in 2010, and TIF 4 in 2012. The state of South Carolina won awards in all rounds of TIF to implement TAP and ultimately established the program in more than 95 schools. Thirty schools adopted TAP in the 2007-2008 academic year, 16 schools in the 2008-2009 academic year, 25 schools in the 2010-2011 academic year and the remaining schools adopted TAP in 2012 and beyond. As discussed in Section 3.2, TAP's staggered adoption in South Carolina forms the basis of our identification strategy.

In addition to implementing the first two TAP elements, South Carolina TAP uses a comprehensive performance pay scheme based on different aspects of the teaching profession reflected in the second two TAP elements. Specifically, 40 percent of teachers' bonus allocation depends on classroom observation scores. Teachers are evaluated at least four times during the school year and a final score is obtained by taking the average of all evaluation scores during the academic year. The other 60 percent is split evenly between individual teacher-value added and school-level value-added scores. Teachers can receive performance bonuses in each of the three categories and must individually achieve higher scores in the first two domains to be eligible for additional awards. For teachers in grades and subjects in which state assessments are not administered, bonus allocation is based on school achievement growth and teacher practices and finally, school administrators can also receive performance pay and these bonuses ranged from \$0 to \$14,000 for the 2009-2010 academic year (South Carolina Department of Education, 2012).[4] For their additional duties, master teachers earn an annual compensation of up to \$15,000, while mentors earn a minimum compensation of \$2,500.

Several comments on the incentive pay scheme under South Carolina TAP are warranted. First, there is no consensus on how to design optimal teacher incentives (Jackson and Bruegmann, 2009; Muralidharan and Sundararaman, 2011; Fryer, 2013; Goodman and Turner, 2013; Imberman and Lovenheim, 2015; Brehm et al., 2017). While it is conceivable that individual incentives dominate group-based incentives because of the free-riding problem inherent in group incentives, complementarities and gains to cooperation may ultimately make group-based incentives a more effective tool. South Carolina TAP is a hybrid program involving both individual and group incentives and thus it is less likely to suffer from the design-specific concerns of simpler incentive pay schemes. Second, bonuses were substantial and sufficiently differentiated to cause changes in the behavior of educators. For example, the average incentive pay for teachers across the state was approximately \$2,000, ranging from \$0 to \$10,000, for the 2009-2010 academic year (South Carolina Department of Education, 2012). As such, the maximum incentive amount was equal to roughly 20 percent of the average annual salary of a public school teacher and it was five times the average bonus pay. Third, incentive pay was not solely determined by teaching outcomes. Teaching practices, coupled with professional feedback, played an equally important role in the award allocation. This is important

[4]The school value-added scores make up 75 percent of the award allocation for school administrators. The remaining 25 percent is based on the program review score measuring the fidelity of TAP implementation in the school.

because the lack of a meaningful feedback due to complex nature of value-added scores is viewed as one potential explanation of why many pay schemes fail to improve student achievement (Fryer, 2013). Finally, while achieving a threshold is sufficient for bonus pay, higher scores enable teachers to extract a larger share from the total available pool.[5] In this respect, the structure of the bonus pay includes both absolute targets and rank-order tournament and does not necessarily imply egalitarian distributions where an overwhelming majority of teachers receive the same award. Appendix B provides details of TAP compensation using a hypothetical example.

3 Data and Methods

In this section, we describe the student, criminal justice, and social welfare program records used for our analysis. We follow this by explaining the empirical methodology behind our findings.

3.1 Data

3.1.1 Data Sources

The data for this study are compiled from several different sources. The first is administrative records from the South Carolina Department of Education (SCDOE). The data include student race, gender, free/reduced lunch status and age, test scores from selected grades and information on high school graduation. In addition, for a subset of academic years, we have records of attendance for each student. Unique identification numbers allow us to track all the students through their tenure in the public school system from the fall of 2000 to the spring of 2017. The SCDOE data do not include information on individual teachers. It is thus not possible to link students to teachers.

The juvenile crime data come from the South Carolina Department of Juvenile Justice (SCDJJ) and include the universe of detailed arrest records from 2000 to 2017. For each juvenile offender file, we have basic demographic information on the arrestees, offense date and the type of crime they are arrested for. We complement these data by drawing information on administrative records from the South Carolina State Law Enforcement Division (SLED) over the same period. Similar to offender files in SCDJJ, adult crime

[5] On average, each TAP school allocates $2,000 to $3,000 per teacher to establish the award pool (Institute of Education Sciences, 2015).

data include demographic information, date of offense and arrests by category of crime.

Finally, we use data from the South Carolina Department of Social Services (SCDSS) for information on enrollment in social programs, which is available through 2019. We are able to link individuals' records across these four data sets. In addition, as part of our mechanism analysis, we rely on publicly available school report cards for data on several school-level attributes, such as measures of school climate, teacher turnover rates, percentage of teachers with advanced degrees, and so on.

Note that because we observe all public school enrollments in the state, concerns about student attrition only arise if students leave the state, attend a private school or are home-schooled. It is possible that students in schools adopting TAP respond by moving to another state or transferring to a private school, but as shown in Section 4.1, timing of TAP implementation is not correlated with the likelihood of attrition from the public education sample. Enrolling in a private school/homeschooling does not generate attrition in our crime and government assistance data because the only relevant margin of attrition in these cases is out-of-state migration.[6]

3.1.2 Sample and Matching Procedure

Our sample consists of first-time eighth graders from the 2002-2003 to 2012-2013 academic years, roughly corresponding to the cohorts born between 1988 and 1999.[7] We choose these particular cohorts primarily because all schools (associated with the first 3 rounds of TIF) adopted TAP between the 2007-2008 and 2010-2011 academic years. Online Appendix Table A.1 shows the implementation years of TAP and associated outcome years.

Given the majority of TAP adopters are high-need schools serving large fractions of disadvantaged students, one would expect TAP schools to be different than the average school in the state. In order to address such differences and to circumvent potential confounding effects, we rely on propensity score matching to identify a set of comparable schools which are most similar to TAP schools in terms of observable characteristics prior to the adoption of TAP (Abadie et al., 2010). In doing so, we estimate a logit model where the dependent variable is an indicator function that takes the value of one if the school

[6] Using the American Community Survey data, we find that less than 7 percent of the population born in South Carolina in 1990s left the state at age 18 or earlier.

[7] We tested for post-adoption sorting into TAP schools and find no evidence of student sorting based on eighth grade cohorts (Section 4.1).

has ever adopted TAP over the sample period and zero otherwise. We select covariates using an adaptive least absolute shrinkage and selection operator (LASSO) procedure, as well as added other school characteristics that we believe should be part of the propensity score model.[8] Online Appendix Table A.2 presents these school characteristics from the baseline academic year.

We estimate the propensity score for being a TAP school and sort the comparison candidates by predicted scores in descending order and select the top 5 percent of non-treated schools. As shown in Column 5 of Online Appendix Table A.2, we fail to reject mean tests of equality for all but one school characteristic. This stands in sharp contrast to the differences in the means between TAP and all other schools in the state whose grade configuration includes eighth grade (Column 6).

Although the matched comparison school sample improves upon the potential comparison sample in terms of alignment with TAP schools, post-matching differences in observable characteristics are not completely eliminated. We believe these discrepancies do not pose a serious threat to identification for at least two reasons. First, our results are not sensitive to the inclusion of (pre-determined) individual and grade-level control variables. Second, as discussed in detail in Section 4.3, the estimated effects of TAP from alternate comparison groups are very similar to those reported throughout the text. Our main alternative comparison group is "future adopter" schools, those schools adopting TAP in 2012 or beyond as part of TIF 4, which have very similar characteristics to pre-adoption TAP schools and for which we fail to reject a test of equality for all characteristics (Online Appendix Table A.2).

3.1.3 Outcomes

Using these unique sources of linked administrative data, we are able to observe several medium- to long-run outcomes for each student in our sample. Measures of educational attainment include twelfth grade

[8] In order to overcome covariate-selections problem in propensity score matching, we use a LASSO of being a TAP school on the following baseline school level controls: the fraction of eighth-grade students who are female, Black, white and free/reduced-price lunch eligible, grade size, school's total enrollment, attendance rate, the fraction of students suspended/expelled, number of full-time teachers, the fraction of teachers with advanced degrees, the fraction of teachers with continuing contracts, teacher turnover rate, the fraction of teachers satisfied with learning environment, the fraction of teachers satisfied with social and physical environment, the fraction of teachers satisfied with home-school relations, the fraction of students satisfied with learning environment, the fraction of students satisfied with social and physical environment, the fraction of students satisfied with home-school relations, the fraction of parents satisfied with learning environment, the fraction of parents satisfied with social and physical environment, and the fraction of parents satisfied with home-school relations. The covariates selected, based on adaptive LASSO procedure, are: the fraction of eighth-grade students who are free/reduced-price lunch eligible, the fraction of students suspended/expelled, the fraction of teachers with continuing contracts, and the fraction of teachers satisfied with social and physical environment. Restricting the propensity score estimation to include only LASSO selected covariates does not change our results in a meaningful way.

enrollment status and on-time high school graduation.[9] Records from the SCDJJ and SLED allow us to examine criminal activity from adolescent to early adulthood. We construct several different measures of crime, including whether or not the student was ever arrested as a juvenile, whether or not the student was ever arrested as an adult between ages 17 and 18, and criminal involvement by severity of crime (felony and non-felony). It is important to note that the upper age of juvenile court jurisdiction over the sample period was 16 and we can measure criminal activity in early adulthood for all cohorts without censoring up to age 18. Online Appendix Table A.1 shows outcome availability by cohort.

Records from the SCDSS allow us to construct a comprehensive measure of economic self-sufficiency: whether or not the student ever received food stamps (renamed Supplemental Nutrition Assistance Program [SNAP] in 2008) or Temporary Assistance for Needy Families (TANF) as an adult between ages 18 and 22. Given most recent cohorts will not be old enough by the end of sample period, our analysis of economic self-sufficiency focuses on earlier eighth grade cohorts (i.e., 2002-2010) and schools adopting TAP as part of TIF 1 and TIF 2.

It is worth mentioning that these conditional cash and in-kind transfers constitute an important source of income for recipients in South Carolina. Using the 2010-2019 SNAP Quality Control files provided by Mathematica Policy Research, Inc., we find the average monthly SNAP benefit ($210 in 2015 dollars) to be roughly equal to 20 percent of the total gross income recipients reported. Finally, to perform a comprehensive evaluation of TAP and explore various mechanisms, we also consider several shorter-run outcomes (e.g., being held back in ninth grade, mandatory high school exit exams taken in the Spring of tenth grade) throughout the paper. The tests and test scales administered in elementary and middle schools changed dramatically beginning with the 2008-2009 academic year which prevent us from analyzing the efficacy of the program on eighth grade test scores. The change was made in an effort to provide a more comprehensive assessment of student learning and ensure that the state's standardized testing program is in line with current educational standards.[10]

[9] Our analysis excludes eighth graders from the 2002-2003 academic year when the outcome of interest is on-time graduation because SCDE provided information on graduation beginning with the 2007-2008 academic year.

[10] The Palmetto Achievement Challenge Test (PACT) was administered to students in select grades since 1999. The South Carolina Palmetto Assessment of State Standards (SCPASS) replaced PACT beginning with the 2008-2009 academic year.

3.1.4 Descriptive Statistics

Columns 1-4 of Table 1 present descriptive statistics for a total of more than 43,000 students from 31 unique schools. Online Appendix Figure A1 shows the distribution of grade configuration for these schools based on the highest grade offered. We show tabulations for the treated sample, by timing of TAP adoption, and for a matched comparison sample. As displayed in Panel A, Black and White students comprise 53 and 43 percent of all students in TAP schools, respectively, and 67 percent of the treated sample received free/reduced-price lunch (Column 1). Students in matched comparison schools are more likely to be Black, come from disadvantaged families, and have lower baseline composite test scores (Column 4).[11] The mean twelfth grade enrollment over the pre-adoption period is 62 percent in TAP schools while it is 67 percent for non-TAP schools (Columns 2 and 4, Panel B). We observe similar differences in criminal justice outcomes between TAP and matched comparison schools. For example, the fraction of individuals who were arrested of a felony crime at 18 or younger is 5.6 and 4.5 percent in these schools, respectively. In contrast, 51 percent of students received government assistance during early adulthood in TAP schools while the rate of reliance on social programs is almost 61 percent in comparison schools.

Finally, the last column shows the same descriptive statistics from our main alternate comparison group, the so-called "future adopters." This alternate sample is very similar to that from Column 2 in observable student characteristics, but the sample size is almost two-thirds of our preferred comparison group. As noted above, the similarity of the estimated impacts of TAP from alternate comparison groups may provide assurance as to the credibility of the identification strategy.

3.2 Empirical Methodology

To evaluate the effects of TAP on student outcomes, we use variation in when and where schools adopted TAP in a difference-in-differences framework and estimate the following equation

$$Y_{isc} = \beta_{DiD} TAP_{sc} + X'_{isc}\Gamma + \delta_s + \lambda_c + \epsilon_{isc} \quad (1)$$

where Y_{isc} is the outcome of interest, e.g., an indicator variable that takes the value one for on-time high

[11] Composite standardized test score is the average of standardized test scores in ELA and math and is available for 33,459 students in our analysis.

school graduation for student i, in school s, and cohort c. The indicator TAP_{sc} is equal to one in the schools and cohorts exposed to TAP, based on eighth grade school. X'_{isc} is a set of observable student and grade composition characteristics, which include birth-year fixed effects and indicators for gender, race, and free/reduced-price lunch status, the fraction of students who are female, Black and free/reduced-price lunch eligible at the school-by-grade level. We also include δ_s and λ_c, which denote school and cohort fixed effects, respectively. Finally, ϵ_{isc} is the error term. Identifying variation comes from two sources: within school differences before and after TAP adoption, and TAP versus non-TAP differences in the same calendar year. Since TAP_{sc} captures different cohorts of student exposed at different times for different lengths of time, β_{DiD} is a weighted average of the overall TAP exposure effect during the outcome years we focus on.

The benefit of the DiD approach is that it increases statistical precision and summarizes impacts over the outcome time horizon, with a single indicator for which it is easy to compare across multiple specifications. However, to investigate dynamic response to treatment, we also estimate flexible event study specifications of the following form:

$$Y_{isc} = \sum_{\substack{\tau=-5 \\ \tau \neq -6^+}}^{4} \gamma_\tau 1(t - t_s^* = \tau) + X'_{isc}\Psi + \delta_s + \lambda_c + \epsilon_{isc} \tag{2}$$

where the TAP indicator is parameterized over time to allow for dynamic treatment effects. The year since TAP adoption is indicated by τ with t_s^* being the year of school-level TAP adoption. Each $1(t-t_s^* = \tau)$ is an indicator variable equal to one for each of the years before and after TAP adoption. The endpoints from the years prior to adoption are combined into an indicator variable for 6 or more years before ($1(t - t_s^* \leq -6)$), and all post-adoption years are displayed. The excluded category are the eighth grade cohorts from 6 or more years before TAP adoption, $\tau = -6^+$, and untreated units are included in this group as well. All other variables are as previously defined. In Section 4.3, we demonstrate that the estimates from the event study model remain robust to alternate choices of the event time window, revealing that binning the pre-endpoints does not pose any threat to identification.

Treatment effects that occur in response to TAP and vary over time are indicated by γ_0 to γ_4 and trace out impacts on student outcomes by cohort of exposure to TAP. For example, in a school with grade 6-8 configuration, students in the initial exposure cohort will be in a TAP school for a single year (eighth grade).

Students in the next cohort are typically exposed to TAP for two years (seventh and eighth grade), and students in the next and subsequent cohorts are exposed for three years (sixth, seventh, and eighth grades).[12] We thus refer to time since treatment indicators as the first through fifth "post-adoption cohort." The event study model also allows us to test for parallel trend condition. The existence of any lag effects (γ_τ for $\tau < 0$)) is likely to invalidate our identification strategy.

Although the lack of large and significant lag effects is assuring in terms of causal interpretation, the two-way fixed effects models can still be susceptible to different forms of biases in settings with staggered treatment adoption (Callaway and Sant'Anna, 2020; Borusyak et al., 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021). More precisely, unless strong assumptions on treatment homogeneity hold, any γ_τ can be expressed as a linear combination of group-specific effects from both its own period and other relative periods. These treatment effects from other relative periods will not cancel out and will contaminate the estimate of γ_τ. In our context, the homogeneity assumption entails early and late TAP adopters experience the same path of treatment effects. This may not be true. For example, treatment effects may vary for early and late TAP adopters because of teachers' mobility, and thus changes in teacher quality, across districts over time. To probe these concerns, we estimate the event study coefficients in equation (2) using the imputation estimator of Borusyak et al. (2021).

The imputation estimator purges this source of bias by generating predicted values of the outcome for students in TAP schools in the post-adoption period using the two-way fixed effects model described above for only the non-treated observations (students in non-TAP schools and yet-to-adopt TAP schools). An estimate of the treatment effect can then be obtained for each treated observation by calculating the difference between their observed and predicted outcome and taking the average of these differences. As a further robustness check, we estimate the event study models using the interaction weighted estimator of Sun and Abraham (2021).[13]

[12]The public schools in South Carolina vary in terms of their grade configuration. There are several primary schools serving students until the end of sixth grade, several other schools contain a grade K-8 configuration and there are also a number of middle schools with a grade 7-9 configuration. This heterogeneity in grade span also highlights the fact that it is not possible to define all students by their sixth grade schools. We exclude eighth grade cohorts immediately preceding the year of TAP implementation in schools with a grade 7-9 configuration. These cohorts are likely to be exposed to TAP for a year in their ninth grade, although keeping them in the analysis sample does not change any of the results. Such schools comprise around 20 percent of all schools in the analysis sample. See also Online Appendix Figure A1.

[13]The Sun and Abraham estimator purges potential biases in settings with staggered treatment adoption by comparing TAP schools only to non-TAP schools and removing yet-to-adopt TAP schools.

Another important threat to identification in settings with variation in treatment timing stems from the negative weighting problem (Borusyak et al., 2021; Goodman-Bacon, 2021). In its simplest form, the issue is related to the weighting scheme implicit in OLS where the weights of the dependent variable are proportionate to the residuals in a regression of treatment on right-hand side variables. The linear probability model can generate fitted values that are greater than one, causing corresponding outcome values to be negatively weighted. This problem is more salient for earlier treated units because fitted values for these units are larger at longer horizons, meaning short-run effects can be over-weighted, while long-run effects are under-weighted. The extent of bias from negative weighting can be severe and may even cause DiD estimates in equation (1) to lie outside the convex hull of the time-varying effects γ_0 to γ_4. As a result, we complement our analysis by estimating a semi-dynamic specification under the assumption of no pre-trends (i.e., γ_τ for $\tau < 0$ are set to zero).

Prior to continuing, it is worth mentioning that this negative weighting problem is inherently different than the preceding source of bias because it arises from the heterogeneity across τ, rather than from the heterogeneity of treatment effects across groups and periods for a given τ. Finally, unless otherwise stated, standard errors are clustered at the school level to allow for dependence in student outcomes within schools.

4 Results

This section reports the results from our analytic strategy, first verifying that our context is not compromised by (i) differential trends between TAP and non-TAP schools, (ii) student sorting, and (iii) attrition in response to the program. We then present our main results on educational attainment, criminal justice, and economic self-sufficiency outcomes, followed by a series of robustness checks that verify our findings.

4.1 Identifying Assumptions, Sorting, and Attrition

The DiD, semi-dynamic models, and event study approaches all rely on the same two assumptions: (i) TAP adoption is not correlated with any prior trend in long-run outcomes across schools, and that (ii) there are no coincident shocks or policy adoptions that could account for the TAP effect. We provide three sets of evidence of the plausibility of the first assumption. First, we test whether TAP adoption was preceded by a systematic change in school characteristics. To diagnose the importance of any pre-existing trend, we

estimate a modified event study by replacing the pre-TAP indicators with a linear trend. The parameter of interest in this specification yields the slope of school characteristics over time prior to TAP adoption. The first column of Table 2 presents these coefficient estimates. Of the 11 outcomes we analyze, none is statistically significant at even the 10% level. Second, we examine the associations between the year of TAP adoption and baseline school characteristics. As shown in Column 2, the covariates do not significantly predict the timing of TAP adoption. The p-value for joint significance is 0.56. The inference on the predictive power of the covariates does not change when we replace the year of adoption with indicator variables for being an early or late adopting school, as reported in Online Appendix Table A.3.

Finally, Figure 1 depicts the cohort-specific point estimates by years elapsed relative to TAP implementation for key student outcomes. The length of the bars extending from each point represents the bounds of the 95% confidence interval. The lagged effects are generally small in magnitude and individually statistically indistinguishable from zero. The pre-adoption coefficients are also jointly equal to zero across all panels.[14] Taken together, we see no trends in educational attainment, crime and self-sufficiency outcomes from cohorts in TAP schools prior to TAP adoption. The second assumption is not directly testable. However, we show in Online Appendix Table A.4 that when we characterize TAP implementation at the district level, with TAP adoption beginning when any school in the district introduces TAP program, we find no impacts on student outcomes. This implies that TAP is not part of some larger package of district-level programming adopted at the same time.

Next, we test for post-adoption student sorting. Table 3 shows the impact of TAP exposure on eighth grade student characteristics. Neither the proportion of girls, Black students, nor students that received free/reduced-price lunch was changed by exposure to TAP, implying that students' families did not switch schools or neighborhoods to access (or avoid) TAP. Similarly, there is little difference in size of the grade cohort or overall school enrollment, nor for prior test scores. As noted in Section 3.1, because the tests and test scales changed dramatically beginning with the 2008-2009 academic year, we limit our analysis in the last column of Table 3 to those students who were enrolled in fifth grade prior to 2008.

Finally, we examine whether TAP adoption is correlated with sample attrition. Differential attrition between TAP and non-TAP schools may lead to a selected sample and, for that matter, may bias the effects

[14]The corresponding p-values in Figure 1 for joint significance are 0.84 in Panel A, 0.33 in Panel B, 0.38 in Panel C, 0.69 in Panel D, 0.63 in Panel E and 0.33 in Panel F.

of the program. To investigate this possibility, we created an indicator variable that takes the value one if the student had not ever enrolled in ninth grade in a South Carolina public school and use it as dependent variable in equation (1).[15] We utilize ninth grade enrollment for the attrition exercise because the state required students to stay in school until 16 over the analysis period and TAP may have a direct effect on dropout over time. The estimated effect of TAP from this analysis is 0.0004 (s.e.=0.0041) which does not suggest any contamination due to attrition (Column 1 of Online Appendix Table A.6).

4.2 TAP and Long-Run Outcomes

We present our baseline DiD results on the relationship between TAP and long-run outcomes in Panel A of Table 4. All estimates include controls for birth year, cohort, and school fixed effects, as well as student and grade composition characteristics. The DiD estimates from a specification without student and grade level controls are reported in Section 4.3.

We begin by showing impacts on educational attainment in Columns 1 and 2 of Table 4. We find that exposure to TAP increased the likelihood of ever being enrolled in twelfth grade by a statistically significant 3.5 percentage points. Taking the mean enrollment of 67.3 percent in non-TAP schools as our benchmark, the estimated impact implies an average increase of 5.2 percent. We analyze the association between TAP and on-time graduation in Column 2 of Table 4. The TAP impact on high school graduation, where data are available for one fewer cohort than twelfth grade enrollment, is 3.8 percentage points, similar in magnitude to the twelfth grade outcome, and statistically significant. This is an average increase of almost 6 percent relative to the comparison mean of 66.4 percent.[16]

We additionally estimate a semi-dynamic model where we allow the effect of TAP to differ depending on time-since-treatment. In doing so, we utilize the imputation estimator of (Borusyak et al., 2021) because dynamic effects, regardless of the relative period, are susceptible to bias resulting from treatment heterogeneity. The benefits of TAP may compound because students are exposed for more of their middle school years and teachers and administrators gain experience with the program. This is exactly what we find, as demonstrated in Panel B of Table 4 and visualized in Panels A and B of Figure 1. For example,

[15] Recall also that attrition in public education occurs if students leave the state or enroll in private school/homeschooling. The only relevant margin in crime and economic self-sufficiency data is out-of-state migration.

[16] Note that the comparison group mean will be lower than published graduation statistics for South Carolina since we count anyone who disappears from the data as if they had not graduated.

the first column indicates that the implementation of TAP increased the probability of ever being enrolled in twelfth grade by a statistically significant 2.1 percentage points for the first post-adoption cohort, while the coefficient estimate for the fourth post-adoption cohort is 7.7 percentage points. The estimated effects for on-time graduation of the same cohorts are 0.030 (s.e.=0.022) and 0.087 (s.e.=0.039), respectively (Column 2). Apart from highlighting the existence of a plausible dose-response relationship for educational attainment, these findings also suggest that negative weighting problem does not bias our findings. As such, the DiD estimates always lie within the convex hull of the time-varying effects reported in Panel B. This is likely be due to the non-trivial size of the comparison group.

Next, we examine the effect of TAP on criminal involvement. Columns 3 and 4 of Table 4 displays the results of this analysis by juvenile (ages 14 to 16) and adult arrests (ages 17 to 18), respectively. The point estimates in Panel A are negative but fall short of statistical significance. Online Appendix Table A.5 presents the same results by disaggregating arrests based on their severity (felony and non-felony offenses). It appears that the impact of TAP on crime is more pronounced (relative to control mean) and precisely estimated for felony offenses committed both in adolescence and early adulthood. Thus to increase power, we further group together juvenile and adult felony crime involvement and define an indicator that takes the value of one if a student was ever arrested of a felony crime at or before age 18. Using this indicator as our outcome of interest, we find that being exposed to TAP decreased the likelihood of ever being arrested of a felony crime at age 18 or earlier by a statistically significant 1.4 percentage points (Column 5 of Table 4). This represents a decrease of 31 percent relative to the control mean. The analogous point estimate for non-felony offenses, reported in the last column of Online Appendix Table A.5, continues to be smaller in magnitude and indistinguishable from zero. Online Appendix Table A.6 also shows the results for types of crimes, including violent crimes, alcohol-drug related crimes, property crimes, and other crimes, respectively (Columns 2-5). The DiD estimates for being arrested of different types of crimes at age 18 or earlier are similar in magnitude across columns.[17]

We present the results on criminal involvement from the semi-dynamic specification in Panel B of Table 4. Panels C-E of Figure 1 display these results graphically. The crime-reducing effects of TAP grow over

[17] Simple assault and battery, possession of drugs, and shoplifting are the most common types of arrests in respective crime categories in Columns 2-4 of Online Appendix Table A.6. Other arrests, reported in Column 5, are a heterogeneous group and include myriad offenses ranging from disorderly conduct to forgery.

time since treatment and they also are more precisely estimated. For example, TAP adoption is associated with a 2.1 percentage point decrease in the likelihood of being arrested of a felony crime by age 18 in the fifth year of the program (Column 5). This is about twice the size of the coefficient estimate obtained in the first year of the program. As with educational attainment, this pattern is consistent with a dose-response explanation — greater exposure to TAP results in greater benefits for students.

Finally, we analyze the relationship between TAP and economic self-sufficiency in early adulthood (ages 18 to 22). Recall that this analysis of later-life outcomes focuses on earlier eighth grade cohorts (2002-2010) and schools adopting TAP through TIF 1 and TIF 2, as more recent cohorts are not yet old enough for us to observe the receipt of government assistance by age 22. As shown in the last column of Table 4, exposed students were, on average, 2.7 percentage points less likely to rely on social welfare programs, which represents a 4.4 percent decrease relative to the comparison mean of 60.7 percent. The influence of TAP on economic self-sufficiency also continues to be more pronounced for cohorts with greater exposure to TAP.

To summarize the long-run results, we also construct an outcome index which is an equally weighted average of the standardized (z-scores by the academic year) measures for our key outcomes (the binary indicators for ever being arrested of a felony offense and reliance on social welfare programs are reverse coded in the construction of the index). The index allows us to obtain an estimate of the overall impact and reduces the chance of false positives due to multiple hypothesis testing (Kling et al., 2007). As shown in the last column of Online Appendix Table A.6, the point estimate from this exercise is 0.047 and indicates that being exposed to TAP is associated with 4.7 percent of a standard deviation increase in outcome index.

To put the estimates in perspective, we compare them to other studies in the related literature. For example, Jackson et al. (2020) find that attending a school with one standard deviation higher predicted test score value added increased (decreased) high school graduation (school-based arrests) by 1.3 (13) percent for ninth grade students in Chicago public schools. The average impact of TAP on graduation is roughly equivalent to attending a school with 4.6 standard deviation higher test score value added, while the estimated impact on felony offenses maps onto attending a school with 2.4 higher standard deviation in test score value added. Cook and Kang (2016) show that delayed school entry eligibility decreased enrollment in twelfth grade by 4 percent in North Carolina. Children born just after the school entry eligibility date

were also 14 percent more involved in serious adult crimes. The effect sizes we obtained here are larger than those of school entry laws. Our estimated effect of TAP on the receipt of social welfare assistance is slightly above half of the food stamp program participation effect resulting from a one percentage point decline in unemployment reported by Currie et al. (2001).

Finally, Online Appendix Table A.7 shows baseline estimates using the interaction weighted estimator outlined in Sun and Abraham (2021). The findings from this alternative approach are consistent with those presented throughout the text. Event studies using the same method further demonstrate that TAP implementation is not correlated with trends in long-run outcomes (Online Appendix Figure A2).

4.3 Robustness Checks and Additional Estimates

We conducted several sensitivity checks to examine the robustness of our results. Since the difference-in-differences estimate nicely summarizes the effect in a single coefficient and the semi-dynamic coefficients that account for heterogeneity in impact over time align with the DiD estimates, we use the DiD estimate on our four key outcomes for these specification tests, displayed in Figure 2 (details on the estimates are in Online Appendix Table A.8). Our first series of sensitivity checks considers the inclusion of early and late adopting cohorts. First, we exclude schools adopting TAP in the 2010-2011 academic year to see whether the estimated effects are attenuated in a meaningful way due to inflow of TAP schools as part of TIF 3 at the end of our sample period ("late adopters"). The DiD estimates are larger in magnitude. Second, our results are also robust to excluding schools adopting TAP as part of TIF 1 in the 2007-2008 academic year ("early adopters").

We then turn to alternative control variables. Third, we examine the sensitivity of our results to conditioning on fifth grade composite standardized test scores. Including this control in the specifications does not affect our estimates though the sample size is smaller due to the availability of baseline scores.[18] Fourth, we estimate a model that interacts several baseline school-level variables (school's total enrollment, attendance rate, the fraction of students suspended/expelled, number of full time teachers, the fraction of teachers with advanced degrees, the fraction of teachers with continuing contracts, and the fraction of teachers satisfied

[18] In a separate exercise to examine threats to validity, we also replace the outcome of interest in equation (1) with fifth grade test scores. Reassuringly, the estimated effect of TAP from this placebo analysis is small in magnitude and statistically indistinguishable from zero; the point estimate is -0.027 (s.e.=0.028). This further confirms lack of sorting into TAP exposure (Column 6 of Online Appendix Table A.6).

with social and physical environment) with a linear trend. In doing so, we allow TAP adoption to be related to different underlying time trends in long-run outcomes across schools, depending on baseline school controls. The point estimates from this extended specification are not different than those presented in Table 4. Fifth, we control for district-specific linear pre-trends, following the two-step procedure proposed by Goodman-Bacon (2021). The estimated effects of TAP are similar to our main model though the confidence intervals are larger. Sixth, we analyze the relationship between TAP and long-run outcomes by excluding student and grade level controls from the specifications. The results indicate that the DiD estimates are not sensitive to the inclusion of control variables, providing further assurance as to the credibility of the identification strategy.

Next we consider the policy environment in South Carolina at the time of the TAP intervention. Seventh, recall that more than 95 schools implemented the program in South Carolina; 10 of which are high schools. Given the feeder structure of public schools (i.e., many middle schools feed into one high school), it is conceivable that some of the students in comparison schools enrolled in TAP high schools and some students in treated schools transitioned to TAP high schools. To assess the contribution of such further exposure, we exclude all students ever enrolled in a TAP high school. The DD estimates are similar to our baseline results. Eighth, as discussed above, the validity of our identifying assumption hinges on the absence of confounding shocks or policy changes that occurred at the same time or just after the introduction of TAP. To our knowledge, GEAR UP — a college preparation program — is the only other education policy that may have coincided with TAP. The state of South Carolina was awarded a six-year GEAR UP grant in 2011 to increase the participation of low-income students in postsecondary education. The program was designed to serve an entire cohort of seventh grade students and followed the cohort through high school. Six schools (3 TAP and 3 non-TAP) in our analysis sample were involved in the GEAR UP program. Conditioning on the GEAR UP status of schools or excluding these schools from the analysis does not change any of our findings.

Perhaps most importantly, we next consider two alternative comparison samples to further gauge the robustness of our results. Our first comparison group is defined by selecting the top 10% of comparison schools based on their propensity scores. The second alternate group comprises future adapters – schools adopting TAP in 2012 (or beyond) as part of TIF 4. The point estimates using these samples are very similar

to those obtained based on our primary matched control group.[19] Event studies for long-run outcomes, using both the top 10% of schools and future adopters as comparison groups, are presented in Online Appendix Figures A3 and A4 and these estimates align with the results in Figure 1.

Tenth, we also explore the intensity of treatment by defining TAP exposure more continuously. To do this, rather than using binary classification, we use total potential years of exposure based on the school's grade configuration as the variable of interest. For example, total potential years of exposure for an eighth grader in the fifth year of the program in a school with grade 5-8 configuration is 4 years. The treatment in this case captures both the extensive and intensive margins. To make the results comparable to earlier results, we rescale years of exposure by dividing by the largest years of potential exposure (5 years). This ensures that treatment dosages vary between 0 and 1 and the coefficients represent the effect on the most heavily treated cohort (a change from 0 to 5). All point estimates from this alternative modeling are statistically significant and evidence of a dose-response relationship is further confirmed.

Our next exercises consider the structure of the standard errors. Eleventh, we cluster the standard errors at the school-by-year level and such alternative clustering does not affect statistical significance. Twelfth, to circumvent concerns over potential contamination in the inference procedure that may arise due to small number of schools, we obtain p-values associated with the test of significance using the wild bootstrap t-procedure clustered at the school level (Cameron et al., 2008). We continue to find coefficient estimates that are statistically significant at the 5% level (shown in Online Appendix Table A.8).

Thirteenth, Online Appendix Figure A5 presents the event study estimates by focusing exclusively on cohorts from five years before and five years after TAP implementation. This restriction in the analysis sample ensures balance in event time and addresses any concerns that may arise due to the binning of pre-endpoints in equation (2). The pre-adoption coefficient estimates from this alternate specification closely resemble those presented in Figure 1.

Finally, in Online Appendix Table A.9, we investigate whether differences in grade configuration across schools confound our results. Specifically, we re-estimate the baseline models by (i) excluding schools with grade configuration above ninth grade and (ii) excluding schools with grade configuration above ninth grade and those with grade configuration below fifth grade. The results are not sensitive to these sample

[19] We dropped the 2012-2013 academic year from the analysis when the comparison group is restricted to future adopters.

restrictions.

In addition to these sensitivity analyses, we performed two placebo tests. Our first placebo exercise shifts the analysis sample back in time to the pre-adoption period and focuses on first-time eighth graders from the 2000-2001 to 2007-2008 academic years. The models are estimated as if treated schools first adopted TAP in 2003 rather than 2007, with schools adopting TAP t years after 2007 as if they adopted in 2003+t.[20] The results from this placebo analysis are reported in Online Appendix Table A.10.[21] As expected, we find no effects of TAP during the pre-TAP period. This indicates that the response is due to TAP and not something about the schools that implemented TAP.

Second, we randomly assign TAP adoption years to schools by drawing dates, without replacement, from the actual pool of program implementation years. We do this for 1,000 sets of placebo adoptions. Figure 3 plots the distribution of point estimates. The vertical red lines in each panel denote the values from Table 4. The location of the true estimates in all panels indicate that the likelihood of finding an impact merely by chance is very low.

Finally, in addition to these robustness and placebo checks, we extend our analysis to see whether there are any differential effects of TAP by gender and race. Online Appendix Table A.11 presents these results. We do not observe strong evidence for heterogeneity in the estimated effects of TAP when the coefficients are benchmarked relative to subgroup-specific control means. However, they are consistently less precisely estimated for White students.

5 Mechanisms

Results from the previous section suggest that TAP significantly improved long-run student outcomes. Using several intermediate outcomes, information on individual school report cards, and evaluations of the school climate from a set of annual surveys given to teachers, students and parents, this section discusses possible mechanisms underlying these improvements.

[20] We limit the analysis to include four cohorts of eighth graders from the first placebo wave to avoid overlapping with the actual post-adoption period, i.e., eighth grade cohorts from 2003 to 2006 for schools adopting TAP in 2003.

[21] As noted above, SCDE provides information on graduation beginning with the 2007-2008 academic year and as a result, we do not have a placebo exercise using on-time graduation as the outcome variable.

5.1 TAP and Intermediate Outcomes

We begin by presenting the impact of TAP on cognitive and non-cognitive high school outcomes in Table 5.[22] It appears that TAP implementation influenced students well before their twelfth grade year, with students 3 percentage points less likely to be retained in ninth grade (Column 1). Online Appendix Figure A.6 presents the results for the event study specification. The pre-adoption coefficient estimates hover at zero and are jointly insignificant (p-value=0.84). After the TAP adoption, however, the effect on retention grows larger with relative time.

To comply with accountability policies, the SCDE mandated that all public school students pass an exit examination to receive a high school diploma. The High School Assessment Program (HSAP) was administered for the first time in the 2005-2006 academic year as the state's high school exit examination and consists of two tests: one in English Language Arts and one in math. Students took HSAP in the spring of tenth grade. We use the composite (average of subject test scores) standardized test scores, although the subject specific estimates of TAP are almost identical to that reported for composite scores. As shown in Column 2, on average, students in TAP schools outperformed those in comparison schools by 0.07 standard deviations. Finally, TAP led to fewer days of absence in tenth grade, although this difference is not statistically significant. In short, exposure to TAP improved students' high school performance throughout their high school trajectory. We also conduct an additional attrition exercise for presence in tenth grade by defining an indicator variable that takes the value of one if the student had not ever enrolled in tenth grade in a South Carolina public school. The estimated effect of TAP from this analysis is -0.016 (s.e.=0.014). This is likely bias our estimates in Columns 2 and 3 of Table 5 towards zero. Given concerns on sample selection and that many students were above the minimum school leaving age by the Spring of tenth grade, we opt out of estimating event study models for tenth grade outcomes.

To quantify the share of the treatment effect that is attributable to improvements in these outcomes, we conduct a mediation analysis (Heckman et al., 2013; Gelbach, 2016) by defining a mechanism specification of the following form:

$$IO^{j}_{isc} = \alpha^{j}_{1} TAP_{sc} + X'_{isc}\alpha_2 + \delta_s + \lambda_c + \epsilon_{isc} \tag{3}$$

[22]The available data for these intermediate outcomes vary. South Carolina eliminated the exit exam in 2015. As a result, tenth grade test score data are available between the 2003-2004 and 2011-2012 eighth grade cohorts. The data on school attendance are available for tenth grade cohorts from the 2006-2007 to 2008-2009 and 2010-2011 to 2014-2015 academic years.

where IO^{j}_{isc} denotes the intermediate outcome j. Next, we consider a modified version of equation (1) by including all mechanism variables:

$$Y_{isc} = \beta_1^{Res} TAP_{sc} + X'_{isc}\beta_2 + \sum_j \theta^j IO^j_{isc} + \delta_s + \lambda_c + \epsilon_{isc} \tag{4}$$

where β_1^{Res} captures the component of the estimated TAP effect that is not explained by improvements in intermediate outcomes which also can be expressed as $\beta_1 = \beta_1^{Res} + \sum_j \alpha_1^j \theta^j$. The validity of this mediation analysis hinges on the unbiasedness of the coefficient estimates θ^j, which is a very generous assumption. With this proviso in mind, for each mechanism variable reported in Table 5, we can compute its explanatory power by $\frac{\alpha_1^j \theta^j}{\beta_1}$. Online Appendix Table A.12 displays the results of the mediation analysis. Odd numbered columns reproduce the baseline estimates using student observations with non-missing data on intermediate outcomes, while even numbered columns report the estimated effects of TAP using equation (4). Online Figure A.7 also plots the fraction of the TAP effect that is due to these intermediate outcomes. For example, the decrease in the probability of being retained in ninth grade explains approximately 15 percent of the on-time graduation treatment effect, tenth grade composite test scores explain up to 14 percent of the same effect and student absenteeism, which is a proxy for non-cognitive ability (Gershenson, 2016; Holbein and Ladd, 2017; Jackson, 2018; Jackson et al., 2020), explains 16 percent of the on-time graduation effect. Averaging across four outcome variables, the explained portion of the estimated effect of TAP is around 47 percent.

5.2 TAP, Teachers and School Culture

Although the predictive power of high school outcomes is non-trivial, our results do not speak to the question of why we observe favorable intermediate outcomes for students in TAP schools. To further explore mechanisms, we consider the following school-level domains which are known to be associated with improvements in student well-being: (i) composition of the teacher workforce, (ii) productivity of incumbent teachers, and (iii) school climate. Table 6 presents evidence on whether TAP led to changes in the composition of the teacher workforce. The total number of teachers in TAP schools remained constant; however, the program increased turnover by around 4 percent relative to the comparison mean (Columns 1

and 2).

Considering the average total number of teachers in comparison schools (the third row in Table 6), this increase in turnover is roughly equivalent to a new hire in TAP schools per year in the post-adoption period which is unlikely to generate the long-run effects observed throughout the paper (e.g., TAP increased the probability of on-time graduation by 3.8 percentage points for the second post-adoption cohort). To shed further light on the role of teacher sorting in explaining our findings, we exclude TAP schools that experienced an average change in teacher turnover rates of more than 5 percentage points (in absolute value) between the pre- and post-adoption periods. The results from this exercise are reported in Online Appendix Table A.13. As shown in the first column, the impact of TAP on teacher turnover is almost equal to zero in magnitude for this sample, while the long-run effects of the program on student outcomes are consistently larger (in absolute value) than those reported in Table 4 and continue to be statistically significant (Columns 2-5). It is also important to note that these schools did not experience any changes in other observed characteristics of the teacher workforce.

Furthermore, the results provide evidence on TAP schools attracting lower quality and less experienced teachers compared to school leavers. This is shown by reductions in the fraction of teachers with advanced degrees, highly qualified teachers (highly qualified status of a teacher is defined using a combination of state certification and subject matter knowledge scores in state assessments) and teachers with continuing contracts (Columns 3-5 of Table 6). To obtain an estimate of TAP on school's overall teacher quality, we additionally created an index by averaging the z-scores of the variables from Columns 3-5. The point estimate for the school's teacher quality index, reported in the last column, is negative and statistically significant at the 10% level. Our emphasis on teacher qualifications, subject-matter expertise, teacher's major, and experience as proxies for teacher quality is largely motivated by existing evidence that finds strong and positive effects of these variables on student achievement (Clotfelter et al., 2007; Angrist and Guryan, 2008; Staiger and Rockoff, 2010; Rockoff et al., 2011). Ideally, one would also like to measure teacher's quality via value-added (Kane and Staiger, 2008; Biasi, 2021). However, as noted in Section 3, the SCDOE data do not include information on teacher-student classroom or grade assignments.

The finding from Columns 3-6 is consistent with recent studies documenting a positive relationship between the receipt of an award and odds of switching to high-performing schools. More precisely, bonus

eligibility provides teachers with a credible signal pertaining to unobservable quality that was previously unavailable in the market. Given information asymmetries increase with tenure, inter-school mobility is more prevalent among experienced teachers post-awards (Bates, 2020; Berlinski and Ramos, 2020). Note also that such compositional changes are likely to attenuate our results. This is precisely what we find in Online Appendix Table A.13 — which shows larger impacts when we exclude high-turnover schools — suggesting that the estimated effects of TAP on student outcomes presented throughout the paper may be lower bounds of the true impacts.

Unfortunately, we do not have individual-level teacher productivity measures to assess whether TAP induced incumbent teachers to exert more effort, say, by altering their behavior and teaching practices. In an attempt to shed some light on increased productivity and changes in school climate, we take advantage of a survey administered annually to teachers, students and parents which is an integral part of the state's accountability system. As part of the survey, all three groups of respondents were asked to report whether they are (i) satisfied with learning environment, (ii) satisfied with social and physical environment, and (iii) satisfied with home and school relationship. Responses were measured on a four-point Likert scale ranging from "disagree" to "agree." School report cards include aggregate information on the fraction of teachers, students, and parents who agree with each of these statements. We also create an overall index, which is defined separately for each group of respondents, by averaging the z-scores of satisfaction measures.

Table 7 presents the results from this analysis. The point estimate for overall parental satisfaction, reported in the first column of Panel A, is positive and statistically significant at the 1% level. Being exposed to TAP is associated with 0.42 of a standard deviation increase in parental satisfaction. The coefficient estimates for the individual components (Columns 2-4) suggest improvements in all domains. Interestingly, the results for parents are not replicated among students. The impact of TAP is negative across all columns (Panel B). As such, students appear to be unhappy with changes put into place at their schools. Lower student satisfaction may be due to students being asked to work harder under the TAP regime. Finally, we find that the fraction of teachers who are satisfied with learning and social and physical environment increased following the adoption of TAP, but the coefficient estimates fall short of statistical significance (Columns 2 and 3). Overall, these results align well with explanations related to changes in school climate as well as increases in the productivity of incumbent teachers.

It is conceivable that TAP exposure also helped students attend better high schools. Increased school quality may impact student outcomes through a variety of channels, such as raising returns to investment in schooling, changes in peer quality, increasing opportunity cost of crime, and so on (Lochner and Moretti, 2004; Cullen et al., 2006; Deming, 2011; Carrell et al., 2018). In order to assess this possibility, in Online Appendix Table A.14 we present estimates that control for ninth-grade school fixed effects. The inclusion of these additional school fixed effects has no appreciable impact on the estimates; therefore, high school sorting is unlikely to be the driving force for our findings. Finally, we investigate the relationship between TAP adoption and changes in school leadership as a potential mechanism underlying our findings in two ways. First, we created an indicator variable that takes the value one if there was a change in the school principal from *t-1* to *t* and use this measure as our outcome of interest in a school-level regression. Next, we controlled for principal's years at the school in our student-level regressions. The results from this analysis are presented in Online Appendix Table A.15. As shown in the first column, we do not find any compelling evidence for post-adoption leadership change. Similarly, DiD estimates in Columns 2-5 are not sensitive to the inclusion of principal's years of experience as an additional control variable.[23]

5.3 Contributions of the Program Components

The program's effectiveness raises the question of which element of the program is most important for generating student success. Recall that TAP is a bundled intervention with four key program elements: (i) multiple career paths, (ii) ongoing applied professional growth, (iii) instruction-focused accountability, and (iv) performance-based compensation. We note that we do not have the empirical setting to directly identify the impact of each TAP element separately, since that would entail variations of the program to be implemented differently in a large number of schools. Instead, we turn to the literature and a comparison to an implementation of TAP in a different context.

A growing body of research casts doubt on the efficacy of teacher professional development programs

[23]We also analyze the incapacitation effect of schooling as a potential explanation for our findings on criminal involvement. That is, whether less time spent in school left more time for crime. In an attempt to shed some light on this pathway, we limit our analysis sample to students who had never enrolled in twelfth grade and re-estimated the program impact on criminal justice outcomes. We acknowledge that dividing the sample based on an endogenous variable is problematic, and therefore caution is warranted in interpreting these results. With this proviso in mind, for those students who dropped out of school without enrolling in twelfth grade, we find that TAP led to a 1 percentage point decline in the probability of ever being arrested of a felony offense between ages 17 and 18. This is not consonant with an incapacitation explanation.

and in-service training to improve teacher and student outcomes; the meta-coefficient for general professional development is a statistically insignificant 0.02 of a standard deviation for math achievement (Garet et al., 2008; Fryer, 2017; Loyalka et al., 2019). These concerns carry over to more innovative forms of professional development such as coaching for teachers (Carneiro et al., 2022). Based on this evidence, we believe that it is very unlikely for the first two elements of the program (multiple career paths and ongoing applied professional growth) to account for the entire impact of TAP on student outcomes. However, there is evidence from other settings that teacher observations and feedback can improve student performance (Taylor and Tyler, 2012; Briole and Maurin, 2022; Taylor, 2023) which aligns with the TAP element of instruction-focused accountability.[24]

In an attempt to shed light on the role and design of incentives, we compare South Carolina TAP with its adoption in the Chicago Public Schools. Specifically, using funds from the TIF, the Chicago school system identified 34 high-need schools (K-8) that agreed to participate in a randomized control trial of TAP over a four-year period from 2007 to 2010. South Carolina and Chicago TAP were identical in their comprehensive structure and shared the same design. In practice, however, individual teacher-value added component of performance-based compensation, originally part of the program, was not implemented in Chicago TAP because the data needed to reliably link students and teachers were not available. Therefore, in addition to classroom observation scores, the Chicago program used group incentives (i.e., school-level value-added) to calculate performance bonuses for teachers (Glazerman and Seifullah, 2012). Other aspects of the incentive pay were quite similar across locations: mentor and master teachers, on average, received an additional $7,000 and $15,000, respectively, for their additional duties, and the average annual pay increase for teachers was around $2,000 in Chicago TAP. The student bodies in both intervention sites were similarly high-need students.[25] Thus, the most notable distinction between the programs is the existence of test-based individual teacher performance pay in South Carolina TAP in concert with group-based incentives, whereas Chicago solely provided group-based incentives.

In their evaluation of the Chicago program, Glazerman and Seifullah (2012) find that Chicago TAP did

[24] Such teacher observations and feedback might be considered professional development. They stand in contrast to workshops and content-focused professional development, which is the main form of professional development evaluated in the surveys above.

[25] While students in Chicago TAP are more likely to be Black compared to those in South Carolina TAP, we are not overly concerned about these demographic differences. This is because, as shown in Online Appendix Table A.11, we do not find strong evidence for heterogeneity in the estimated effects of TAP by race.

not lead to improvements in student achievement. This lack of improvement holds true for both lower and upper grades throughout the entire program, contrasting with our South Carolina evidence showing benefits on short-term academic outcomes. Short-term program benefits were also found at other TAP sites which fully implemented the program including individual incentives (Springer et al., 2014; Chiang et al., 2015; Eren, 2019). While we recognize that the discrepancy in these results may not be solely attributable to (unintended) omission of individual incentives, this comparison implies that tying compensation to teacher-value added scores in South Carolina TAP was a crucial component of program success. However we note that individual incentives on their own may not be sufficient for improving student outcomes. Existing evidence regarding the impact of a stand-alone pay scheme focusing exclusively on test-based individual incentives in the U.S. (such as the program studied in Springer et al. (2010)) *also* does not boost student outcomes. Thus we conclude that TAP's effectiveness may come from complementarities within the program itself: individual incentives are necessary but not sufficient.

6 Benefit-Cost Analysis of TAP

In this section, we provide a simple back-of-the envelope cost calculation to put these estimated impacts into monetary perspective. Before proceeding, it is important to keep in mind that any benefit-cost analysis is speculative and subject to several caveats. The total average cost of TAP implementation is about $250 per student (Institute of Education Sciences, 2015). We break the benefits associated with TAP into two components: (i) broader benefits to society originating from reduced crime and (ii) future gains due to increased high school graduation. Recent research suggests that receipt of government assistance — a cost to taxpayers — leads to a wide range of positive outcomes, including improved adult health, better birth and child outcomes and lower criminal involvement (Almond et al., 2011; Hoynes et al., 2016; Tuttle, 2019). Because of this uncertainty in net social gains, we opt out of including the benefit to taxpayers resulting from reduced reliance on social welfare programs. All monetary values are presented in 2015 dollars. We use the marginal value of public funds, which compares recipients willingness to pay for the program to the cost to the government of funding the program, to put these numbers in a single framework (Hendren and Sprung-Keyser, 2020).

We monetize the broader cost of crime by assigning each type of crime the social cost estimates reported

in Miller (1996). These estimates were based on jury award data and we use per victim cost values. For each individual in our analysis sample, we obtain an overall social cost of crime by summing victim cost values from all arrests up to age 18.[26] We use this total cost measure as variable of interest in equation (1) and estimate the impact of TAP on social benefits resulting from averted crimes.[27] Panel A of Table 8 reports the point estimates from this exercise for any criminal activity and felony offenses in rows 1 and 2, respectively. The estimated benefits from reduced crime outweigh the cost of TAP by more than 6 to 1.

Next, we follow Heller et al. (2017) in our calculations of future monetary gains due to increased high school graduation. We assume that each graduate accrues one additional year of education relative to each non-graduate and focus on values related to earnings and health. To estimate the gains associated with earnings, we use synthetic work-life estimates from Julian and Kominski (2011). Work-life earnings represented expected earnings over a 40-year period for the population aged 25 to 64. We take the synthetic lifetime earnings values and divide them by 40 to assign an annual earnings value for each year. Note that this exercise ignores the curvature of the age-earnings profile. We then discount annual earnings at 3 percent to calculate the present value of lifetime earnings of a high school dropout. Assuming a 12 percent increase in lifetime earnings from an additional year of schooling, we calculate the total earnings gain. Education impacts lives beyond earnings. For health returns to education, Cutler and Lleras-Muney (2006) reported a present value between $13,500 and $44,000 in terms of increased life expectancy. We monetize the median value of these estimates as health returns to education. We use the sum of earnings and worth of health resulting from an additional year of education as our measure of graduation benefits, then multiply benefits by an indicator for whether the individual enrolled in twelfth grade and use the result as the outcome of interest in equation (1). Finally, the cost of an extra year of schooling in South Carolina ($346), which we proxy by expenditures per pupil from the Common Core of Data averaged over 2006-2016 ($9,932) again multiplied by an indicator for enrollment in twelfth grade. Our estimate for the net future benefits of graduation from Panel B is around $2,037. Combining the benefits from reduced felony offenses and increased graduation results in a MVPF of 14, making TAP a very cost-effective program.

[26] Our benefit-cost analysis does not take into account direct cost of crime to the justice system. The results thus can be considered a lower bound estimate of the total cost. Additionally, the statistical value of life adds a very high cost to a very small number of fatal crimes. To be more conservative in our estimated benefits, we divide the cost of homicides reported in Miller (1996) by half (Kling et al., 2005; Heller et al., 2017).

[27] We assign negative numbers to the dollar values so that the positive point estimates in Panel A of Table 8 reflect the benefits of reduced crime.

7 Conclusion

Difference-and-differences and dynamic model estimates of the impact of a comprehensive performance pay program for teachers show that it improved longer-term educational attainment and reduced criminal activity and dependence on government assistance for young people exposed to the program. We find little scope for student sorting or changing teacher composition to explain the program effects, and benefits of the program far exceeded its costs. Our analysis also reveals that TAP led to improvements in both students' test-score and non-test-score outcomes throughout their high school trajectory. Finally, using evaluations from a set of annual surveys, we show that teachers and parents both felt more satisfied with the post-adoption learning environment. Taken together, our analysis provides evidence that comprehensive performance-pay programs can be an effective school improvement strategy and help to narrow existing disparities for disadvantaged children. Additionally, limiting evaluation outcomes to shorter-run outcomes may underestimate program effects.

Based on these findings, a natural question to ask is why TAP succeeded when many other U.S. based teacher incentive pay programs failed to improve student outcomes, at least in the short-run. The hybrid nature of incentive design (individual and group incentives), substantial and sufficiently differentiated structure of awards (absolute targets and rank-order tournament), the existence of multiple performance metrics (teaching practices and teaching outcomes) and observation and feedback mechanism may each have contributed to the efficacy of TAP. Our holistic comparison of South Carolina TAP to a version of the program that excluded individual-level incentives implies that individual incentives are a necessary component of the program, however literature on standalone individual incentive program shows no impact on student success. Thus we conclude that TAP's success may be due not any individual component but to the combination of components included in the program.

References

Abadie, A., A. Diamond, and J. Hainmueller (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), 493–505.

Abdulkadiroğlu, A., J. D. Angrist, P. D. Hull, and P. A. Pathak (2016). Charters without lotteries: Testing takeovers in New Orleans and Boston. *American Economic Review 106*(7), 1878–1920.

Almond, D., H. W. Hoynes, and D. W. Schanzenbach (2011). Inside the War on Poverty: The Impact of Food Stamps on Birth Outcomes. *The Review of Economics and Statistics 93*(2), 387–403.

Anders, J., A. C. Barr, and A. A. Smith (2023, February). The effect of early childhood education on adult criminality: Evidence from the 1960s through 1990s. *American Economic Journal: Economic Policy 15*(1), 37–69.

Angrist, J. D. and J. Guryan (2008). Does teacher testing raise teacher quality? evidence from state certification requirements. *Economics of Education Review 27*(5), 483–503.

Atkinson, A., S. Burgess, B. Croxson, P. Gregg, C. Propper, H. Slater, and D. Wilson (2009). Evaluating the impact of performance-related pay for teachers in England. *Labour Economics 16*(3), 251–261.

Bailey, D., G. J. Duncan, C. L. Odgers, and W. Yu (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of research on educational effectiveness 10*(1), 7–39.

Bailey, D. H., G. J. Duncan, F. Cunha, B. R. Foorman, and D. S. Yeager (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest 21*(2), 55–97.

Bates, M. (2020). Public and private employer learning: Evidence from the adoption of teacher value added. *Journal of Labor Economics 38*(2), 375–420.

Berlinski, S. and A. Ramos (2020). Teacher mobility and merit pay: Evidence from a voluntary public award program. *Journal of Public Economics 186*, 104186.

Beuermann, D. W. and C. K. Jackson (2022). The short-and long-run effects of attending the schools that parents prefer. *Journal of Human Resources 57*(3), 725–746.

Biasi, B. (2021, August). The labor market for teachers under different pay schemes. *American Economic Journal: Economic Policy 13*(3), 63–102.

Bleiberg, J., E. Brunner, E. Harbatkin, M. A. Kraft, and M. G. Springer (2023, March). Taking teacher evaluation to scale: The effect of state reforms on achievement and attainment. Working Paper 30995, National Bureau of Economic Research.

Bonilla, S. and T. S. Dee (2020). The effects of school reform under nclb waivers: Evidence from focus schools in kentucky. *Education Finance and Policy 15*(1), 75–103.

Borman, G. D., G. M. Hewes, L. T. Overman, and S. Brown (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of educational research 73*(2), 125–230.

Borman, G. D., R. E. Slavin, A. C. Cheung, A. M. Chamberlain, N. A. Madden, and B. Chambers (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal 44*(3), 701–731.

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:2108.12419*.

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

Briole, S. and É. Maurin (2022). There's always room for improvement: The persistent benefits of a large-scale teacher evaluation system. *Journal of Human Resources*, 1220–11370.

Callaway, B. and P. H. Sant'Anna (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The review of economics and statistics 90*(3), 414–427.

Carneiro, P., Y. Cruz-Aguayo, R. Intriago, J. Ponce, N. Schady, and S. Schodt (2022). When promising interventions fail: Personalized coaching for teachers in a middle-income country. *Journal of Public Economics Plus 3*, 100012.

Carrell, S. E., M. Hoekstra, and E. Kuka (2018). The long-run effects of disruptive peers. *American Economic Review 108*(11), 3377–3415.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly journal of economics 126*(4), 1593–1660.

Chiang, H., A. Wellington, K. Hallgren, C. Speroni, M. Herrmann, S. Glazerman, and J. 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*.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review 26*(6), 673–682. Economics of Education: Major Contributions and Future Directions - The Dijon Papers.

Cohodes, S. R. (2020). The long-run impacts of specialized programming for high-achieving students. *American Economic Journal: Economic Policy 12*(1), 127–66.

Cohodes, S. R., D. S. Grossman, S. A. Kleiner, and M. F. Lovenheim (2016). The effect of child health insurance access on schooling: Evidence from public insurance expansions. *Journal of Human Resources 51*(3), 727–759.

Cook, P. J. and S. Kang (2016). Birthdays, schooling, and crime: Regression-discontinuity analysis of school performance, delinquency, dropout, and crime initiation. *American Economic Journal: Applied Economics 8*(1), 33–57.

Cullen, J. B., B. A. Jacob, and S. Levitt (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica 74*(5), 1191–1230.

Currie, J., J. Grogger, G. Burtless, and R. F. Schoeni (2001). Explaining recent declines in food stamp program participation. *Brookings-Wharton Papers on Urban Affairs*, 203–244.

Cutler, D. M. and A. Lleras-Muney (2006, July). Education and health: Evaluating theories and evidence. Working Paper 12352, National Bureau of Economic Research.

Dee, T. S. and B. Jacob (2011). The impact of no child left behind on student achievement. *Journal of Policy Analysis and management 30*(3), 418–446.

Dee, T. S. and J. Wyckoff (2015). Incentives, selection, and teacher performance: Evidence from impact. *Journal of Policy Analysis and Management 34*(2), 267–297.

Deming, D. J. (2011). Better schools, less crime? *The Quarterly Journal of Economics 126*(4), 2063–2115.

Deming, D. J., S. Cohodes, J. Jennings, and C. Jencks (2016). School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics 98*(5), 848–862.

Deming, D. J., J. S. Hastings, T. J. Kane, and D. O. Staiger (2014). School choice, school quality, and postsecondary attainment. *American Economic Review 104*(3), 991–1013.

Dynarski, S., J. Hyman, and D. W. Schanzenbach (2013). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. *Journal of Policy Analysis and Management 32*(4), 692–717.

Eren, O. (2019). Teacher incentives and student achievement: Evidence from an Advancement Program. *Journal of Policy Analysis and Management 38*(4), 867–890.

Figlio, D. and L. Kenny (2007, June). Individual teacher incentives and student performance. *Journal of Public Economics 91*(5-6), 901–914.

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics 31*(2), 373–407.

Fryer, R. G. (2014). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *The Quarterly Journal of Economics 129*(3), 1355–1407.

Fryer, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments*, Volume 2, pp. 95–322. Elsevier.

Garet, M. S., S. Cronen, M. Eaton, A. Kurki, M. Ludwig, W. Jones, K. Uekawa, A. Falk, H. S. Bloom, F. Doolittle, et al. (2008). The impact of two professional development interventions on early reading instruction and achievement. ncee 2008-4030. *National Center for Education Evaluation and Regional Assistance*.

Gelbach, J. B. (2016). Can simple mechanism design results be used to implement the proportionality standard in discovery? *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 200–221.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy 11*(2), 125–149.

Glazerman, S. and A. Seifullah (2012). An evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after four years. final report. *Mathematica Policy Research, Inc.*.

Glewwe, P., N. Ilias, and M. Kremer (2010, July). Teacher incentives. *American Economic Journal: Applied Economics 2*(3), 205–27.

Goodman, S. F. and L. J. Turner (2013). The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program. *Journal of Labor Economics 31*(2), 409–420.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Gray-Lobe, G., P. A. Pathak, and C. R. Walters (2021). The long-term effects of universal preschool in boston. Technical report, National Bureau of Economic Research.

Hanushek, E. A., J. Luo, A. J. Morgan, M. Nguyen, B. Ost, S. G. Rivkin, and A. Shakeel (2023, March). The effects of comprehensive educator evaluation and pay reform on achievement. Working Paper 31073, National Bureau of Economic Research.

Heckman, J., R. Pinto, and P. Savelyev (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–86.

Heller, S. B., A. K. Shah, J. Guryan, J. Ludwig, S. Mullainathan, and H. A. Pollack (2017). Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago. *The Quarterly Journal of Economics 132*(1), 1–54.

Hendren, N. and B. Sprung-Keyser (2020). A unified welfare analysis of government policies. *Quarterly Journal of Economics 135*(3), 1209–1318.

Holbein, J. B. and H. F. Ladd (2017). Accountability pressure: Regression discontinuity estimates of how No Child Left Behind influenced student behavior. *Economics of Education Review 58*, 55–67.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics 13*(2), 324–340.

Holmstrom, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org. 7*, 24.

Hoynes, H., D. W. Schanzenbach, and D. Almond (2016). Long-run impacts of childhood access to the safety net. *American Economic Review 106*(4), 903–34.

Imberman, S. A. (2015). How effective are financial incentives for teachers? *IZA World of Labor 158*.

Imberman, S. A. and M. F. Lovenheim (2015). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics 97*(2), 364–386.

Institute of Education Sciences (2015). Teacher training, evaluation, and compensation intervention report: TAP: The system for teacher and student advancement.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

Jackson, C. K. and E. Bruegmann (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics 1*(4), 85–108.

Jackson, C. K., S. C. Porter, J. Q. Easton, A. Blanchard, and S. Kiguel (2020). School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights 2*(4), 491–508.

Julian, T. and R. Kominski (2011). Education and synthetic work-life earnings estimates. American Community Survey Reports. acs-14. *US Census Bureau*.

Kane, T. J. and D. O. Staiger (2008, December). Estimating teacher impacts on student achievement: An experimental evaluation. Working Paper 14607, National Bureau of Economic Research.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental analysis of neighborhood effects. *Econometrica 75*(1), 83–119.

Kling, J. R., J. Ludwig, and L. F. Katz (2005). Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment. *The Quarterly Journal of Economics 120*(1), 87–130.

Lavy, V. (2020). Teachers' Pay for Performance in the Long-Run: The Dynamic Pattern of Treatment Effects on Students' Educational and Labour Market Outcomes in Adulthood. *The Review of Economic Studies 87*(5), 2322–2355.

Lochner, L. and E. Moretti (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American economic review 94*(1), 155–189.

Loyalka, P., A. Popova, G. Li, and Z. Shi (2019). Does teacher training actually work? Evidence from a large-scale randomized evaluation of a national teacher training program. *American Economic Journal: Applied Economics 11*(3), 128–54.

Ludwig, J. and D. L. Miller (2007). Does head start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly journal of economics 122*(1), 159–208.

McElroy, K. (2023). Does test-based accountability improve more than just test scores? *Economics of Education Review 94*, 102381.

Miller, T. R. (1996). *Victim costs and consequences: A new look*. US Department of Justice, Office of Justice Programs.

Morgan, A. J., M. Nguyen, E. A. Hanushek, B. Ost, and S. G. Rivkin (2023, March). Attracting and retaining highly effective educators in hard-to-staff schools. Working Paper 31051, National Bureau of Economic Research.

Muralidharan, K. and V. Sundararaman (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy 119*(1), 39–77.

Rockoff, J. E., B. A. Jacob, T. J. Kane, and D. O. Staiger (2011, 01). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy 6*(1), 43–74.

Schueler, B. E., C. A. Asher, K. E. Larned, S. Mehrotra, and C. Pollard (2022). Improving low-performing schools: A meta-analysis of impact evaluation studies. *American Educational Research Journal 59*(5), 975–1010.

Schueler, B. E., J. S. Goodman, and D. J. Deming (2017). Can states take over and turn around school districts? Evidence from Lawrence, Massachusetts. *Educational Evaluation and Policy Analysis 39*(2), 311–332.

Sojourner, A. J., E. Mykerezi, and K. L. West (2014). Teacher pay reform and productivity panel data evidence from adoptions of Q-Comp in Minnesota. *Journal of Human Resources 49*(4), 945–981.

South Carolina Department of Education (2012). The system for teacher and student advancement.

Speroni, C., A. Wellington, P. Burkander, H. Chiang, M. Herrmann, and K. Hallgren (2020). Do educator performance incentives help students? evidence from the teacher incentive fund national evaluation. *Journal of Labor Economics 38*(3), 843–872.

Springer, M. G., D. Ballou, and A. Peng (2014). Estimated effect of the Teacher Advancement Program on student test score gains. *Education Finance and Policy 9*(2), 193–230.

Springer, M. G., L. Hamilton, D. F. McCaffrey, D. Ballou, V.-N. Le, M. Pepper, J. Lockwood, and B. M. Stecher (2010). Teacher pay for performance: Experimental evidence from the project on incentives in teaching. *National Center on Performance Incentives*.

Staiger, D. O. and J. E. Rockoff (2010, September). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives 24*(3), 97–118.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

Taylor, E. S. (2023). Teacher evaluation and training. In E. Hanushek, S. Machin, and L. Woessman (Eds.), *The Handbook of the Economics of Education*, Volume 7. Elsevier.

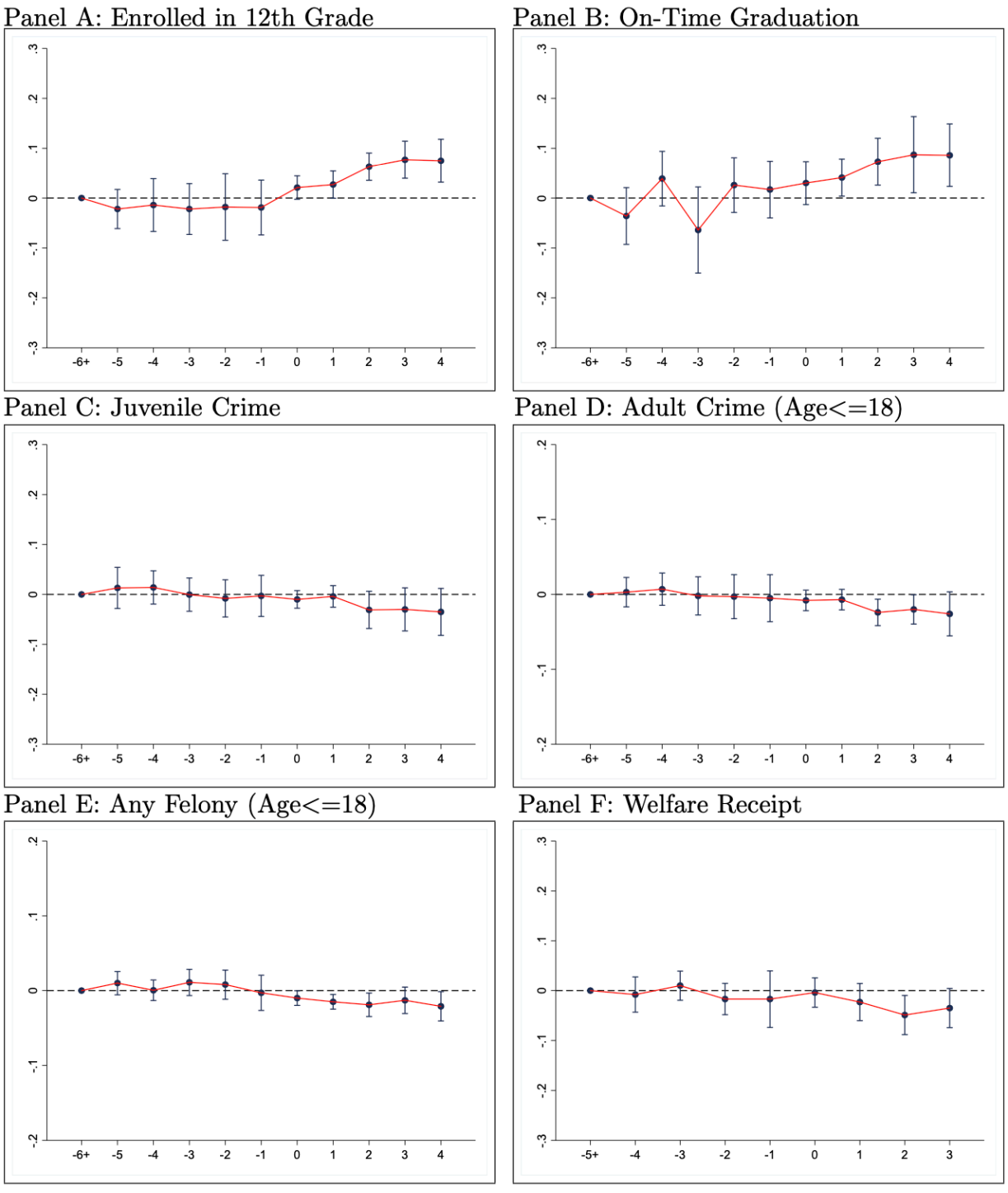
Taylor, E. S. and J. H. Tyler (2012, December). The effect of evaluation on teacher performance. *American Economic Review 102*(7), 3628–51.

Tuttle, C. (2019). Snapping back: Food stamp bans and criminal recidivism. *American Economic Journal: Economic Policy 11*(2), 301–27.

Zimmer, R., G. T. Henry, and A. Kho (2017). The effects of school turnaround in Tennessee's achievement school district and innovation zones. *Educational Evaluation and Policy Analysis 39*(4), 670–696.

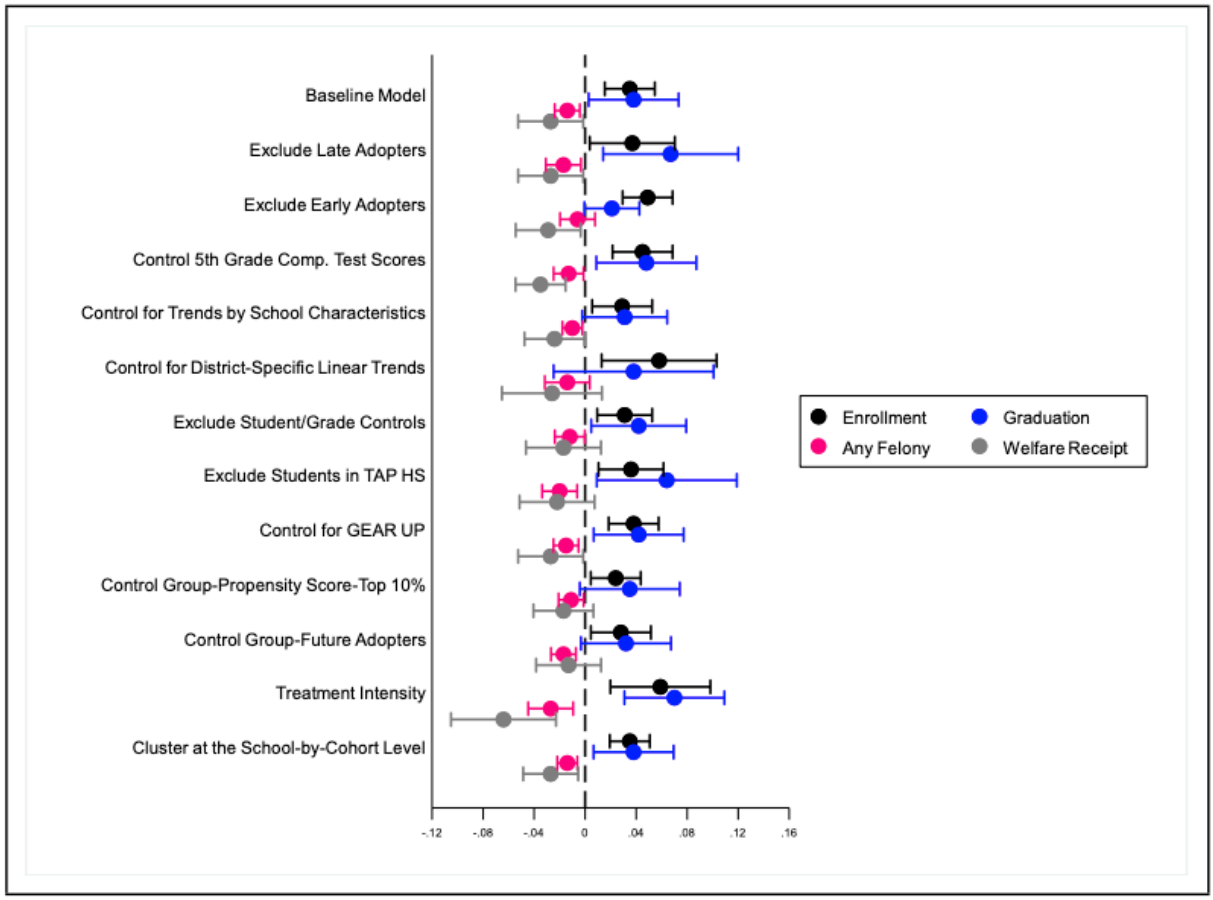
Tables and Figures

Figure 1: Event Study Estimates of the Effect of TAP on Long-Run Outcomes



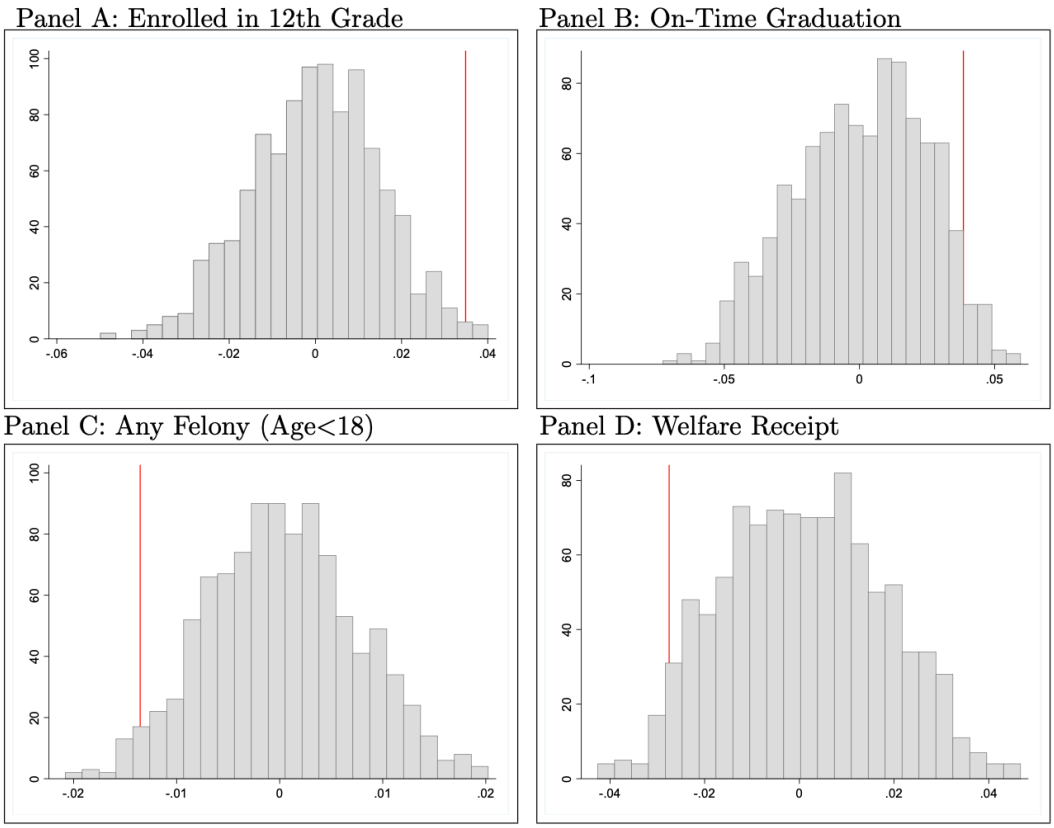
Notes: This figure shows event study estimates for various outcomes obtained using the imputation estimator from Borusyak et al. (2022). Each panel shows coefficient estimates and 95% confidence intervals based on standard errors clustered at the school level. Six or more years before TAP adoption ($\tau \leq -6$) is the omitted category.

Figure 2: Effect of TAP on Long-Run Outcomes: Robustness Checks



Notes: This figure shows various robustness checks for four main outcomes. Difference-in-differences estimates for models that exclude early and late TAP adopters, control for fifth grade composite test scores, control for trends by baseline school characteristics, control for district-specific linear trends, exclude student and grade level controls, exclude students enrolled in TAP high schools, control for GEAR UP status of schools, use alternative control groups and estimate TAP effects more continuously. Each row also shows 95% confidence interval based on standard errors clustered at the school level in Rows 1-12, while the last row displays confidence interval using standard errors clustered at the school-by-year level.

Figure 3: Effect of TAP on Long-Run Outcomes: Placebo Estimates



Notes: This figure shows the distribution of the coefficient estimates resulting from 1,000 sets of random assignments of schools to TAP adoption. The vertical lines denote the actual estimates.

Table 1: **Summary Statistics**

	TAP Schools			Comparison Schools	Alt. Comparison Future Adopters
	All Years	Pre-Adoption	Post-Adoption	All Years	All Years
	(1)	(2)	(3)	(4)	(5)
Panel A: Student Characteristics					
Black	0.528	0.525	0.534	0.726	0.603
White	0.429	0.445	0.405	0.243	0.349
Female	0.491	0.495	0.486	0.491	0.495
Free/Reduced Lunch	0.666	0.648	0.694	0.801	0.693
Baseline Composite Test Scores	-0.466	-0.483	-0.450	-0.537	-0.523
Panel B: Juvenile/Adult Outcomes					
Enrolled in 12th Grade	0.642	0.620	0.675	0.673	0.688
Graduated HS in 4 Years	0.624	0.574	0.686	0.664	0.674
Juvenile Arrest (up to age 17)	0.150	0.159	0.136	0.139	0.123
Adult Arrest (17-18 yrs. old)	0.086	0.095	0.072	0.068	0.066
Any Felony ($\leq$18 yrs.old)	0.056	0.062	0.046	0.045	0.040
Welfare Receipt (18-22 yrs old)	0.514	0.527	0.498	0.607	0.548
Sample Size	29,645	17,761	11,884	13,417	9,575

Notes: This table reports baseline and outcome variables for relevant study populations. The tabulations reflect our research sample which comprises students enrolled in eighth grade for the first time between the 2002-2003 and 2012-2013 academic years. The matched comparison sample in Column (4) is constructed by selecting from all schools in the state a set where baseline student/school characteristics are most similar to TAP schools. Future adopters in Column (5) are schools adopting TAP post-2012. Baseline composite test score is the average of the standardized test scores in English Language Arts and math from fifth grade. Test scores are standardized against the statewide mean and standard deviation by test year-subject.

Table 2: **Trends in School Characteristics Prior to TAP Adoption and Predicting TAP Adoption Year**

		Dep Var:
	Trend	TAP Adoption Year
	(1)	(2)
Fraction of Female Students (8th Grade)	0.003	1,908.03
	(0.004)	(5,354.96)
Fraction of Black Students (8th Grade)	0.006	-763.01
	(0.005)	(1,458.15)
Fraction of Free/Reduced Lunch Students (8th Grade)	0.006	-2,388.68
	(0.005)	(2,594.28)
Total School Enrollment	-1.620	13.119
	(6.716)	(244.38)
Student Attendance Rate (%)	-0.011	24.949
	(0.146)	(30.549)
Percent of Students Suspended/Expelled	0.070	1.652
	(0.854)	(3.802)
Total Number of Teachers in the School	-0.454	-31.505
	(0.429)	(62.424)
Percent of Teachers with an Advanced Degree	1.857	3.635
	(1.571)	(15.761)
Percent of Continuing Contract Teachers	0.655	-8.421
	(0.621)	(19.813)
Percent of Teachers Satisfied with Social and Physical Environment	-0.683	-12.791
	(0.945)	(21.290)
Baseline (5th Grade) Composite Score	0.004	-430.39
	(0.013)	(2,291.37)
F-test (p-value)		0.56
Sample Size	302	31

Notes: Each cell in Column (1) presents a separate regression where the key coefficient of interest is on a trend in the number of years since TAP adoption. The regression specifications, which control for cohort and school fixed effects, include indicators for each post-adoption year and therefore, the point estimates in Column (1) can be interpreted as a test for whether there is a significant pre-trend for each outcome. Column (2) tests whether the year of TAP adoption is associated with school characteristics from the baseline (2002-2003) academic year. Standard errors are clustered at the school level in Column (1), while heteroskedasticity-robust standard errors are reported in Column (2). The F-test p-value comes from a test that the coefficients shown are jointly equal to zero.

Table 3: **Effect of TAP on Student Sorting**

	Fraction of Students (8th Grade)					
	Female	Black	Free/Reduced Lunch	Grade Size (8th Grade)	Total School Enrollment	5th Grade Test Scores
	(1)	(2)	(3)	(4)	(5)	(6)
TAP	-0.005	0.025	0.028	2.800	-32.181	-0.050
	(0.014)	(0.018)	(0.018)	(8.262)	(31.010)	(0.035)
Sample Size	302	302	302	302	302	230

Notes: This table reports difference-and-difference estimates of TAP exposure on the school characteristics listed in the column heading. The effective sample in Column (6) is restricted to students who were in 5th grade prior to the 2008-2009 academic year to account for changes in tests and test scales. The specifications control for school and cohort fixed effects. All outcomes are measured at the school-by-cohort level. Standard errors are clustered at the school level.

Table 4: **Effect of TAP on Long-Run Outcomes**

	Enrolled in 12th Grade (1)	Graduated from HS in 4 Years (2)	Any Juvenile Crime (3)	Any Adult Crime (4)	Any Felony (Age≤18) (5)	Welfare Receipt (6)
Panel A: Difference-in-Differences Estimates						
TAP	0.035***	0.038**	-0.012	-0.011	-0.014**	-0.027**
	(0.010)	(0.018)	(0.012)	(0.007)	(0.005)	(0.013)
Panel B: Semi-Dynamic Model Estimates						
1st Postadoption Cohort	0.021*	0.030	-0.010	-0.008	-0.010**	-0.004
	(0.012)	(0.022)	(0.009)	(0.007)	(0.005)	(0.015)
2nd Postadoption Cohort	0.027*	0.041**	-0.004	-0.007	-0.015***	-0.023
	(0.014)	(0.019)	(0.011)	(0.007)	(0.005)	(0.018)
3rd Postadoption Cohort	0.063***	0.073***	-0.031*	-0.024***	-0.019**	-0.049**
	(0.014)	(0.024)	(0.019)	(0.009)	(0.007)	(0.020)
4th Postadoption Cohort	0.077***	0.087**	-0.030	-0.020**	-0.013	-0.035*
	(0.019)	(0.039)	(0.022)	(0.010)	(0.009)	(0.020)
5th Postadoption Cohort	0.075***	0.086***	-0.035	-0.026*	-0.021**	-
	(0.022)	(0.032)	(0.024)	(0.015)	(0.010)	
Comparison Mean	0.673	0.664	0.139	0.068	0.045	0.607
Sample Size	43,062	38,253	43,062	43,062	43,062	30,081
Controls:						
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports difference-in-differences and semi-dynamic model estimates of the effect of TAP exposure on long-run outcomes. The coefficient estimates in Panel B are obtained using imputation estimator from Borusyak et al. (2022).. Standard errors are clustered at the school level. All specifications control for birth year, cohort, and school fixed effects. Student characteristics include indicators for gender, race, and free/reduced lunch status. Grade composition measures include fraction of students who are female, black, and free/reduced lunch eligible at the school-by-grade level. The dependent variable in Column 1 takes the value one if student was ever enrolled in 12th grade and it takes the value one if student graduated from high school in 4 years in Column 2. The dependent variable in Columns (3)-(5) takes the value one if student was ever arrested as a juvenile or adult. In the last column, the dependent variable takes the value one if student was ever enrolled in social programs (SNAP and TANF) as an adult between ages 18 and 22. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 5: **Mechanisms: Effect of TAP on High School Grade Retention, Test Scores and Student Absenteeism**

	Grade Retention (9th Grade) (1)	Composite Test Score (10th Grade) (2)	Absenteeism (10th Grade) (3)
TAP	-0.029*	0.069**	-2.364
	(0.017)	(0.029)	(2.185)
Comparison Mean	0.120	-0.102	18.22
Sample Size	40,800	27,096	26,323
Controls:			
Cohort and School Fixed Effects	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes

Notes: This table reports difference-in-differences estimates of the effect of TAP exposure on student high school outcomes. The tenth grade test score data are available between the 2003-2004 and 2011-2012 eighth grade cohorts. The data on school attendance are available for tenth grade cohorts from the 2006-2007 to 2008-2009 and 2010-2011 to 2014-2015 academic years. Composite standardized test score is the average of the standardized tests in English Language Arts and math. Standard errors are clustered at the school level. See notes to Table 4 and the text for further details. * significant at 10%, ** significant at 5%.

Table 6: **Mechanisms: Effect of TAP on 8th Grade School Characteristics**

	Total Number of Teachers School	% of Teachers Returning from Previous Year	% of Teachers with Advanced Degrees	% of Highly Qualified Teachers	% of Continuing Contract Teachers (Tenured)	Teacher Quality Index
	(1)	(2)	(3)	(4)	(5)	(6)
TAP	-0.389	-3.481**	-2.309	-2.388	-3.870*	-0.274*
	(1.775)	(1.721)	(2.690)	(1.603)	(2.343)	(0.149)
Comparison Mean	31.64	82.21	54.48	75.72	74.10	0.00
Sample Size	302	302	302	272	272	242
Controls:						
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
School Composition	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports difference-in-differences estimates of the effect of TAP exposure on 8th grade school characteristics. All specifications control for fraction of students who are female, black and free/reduced lunch eligible at the school-by-year level. Standard errors are clustered at the school level. * significant at 10%, **significant at 5%.

Table 7: **Mechanisms: Effect of TAP on Teacher, Student and Parent Satisfaction**

	Overall Satisfaction Index (1)	% Satisfied with Learning Env. (2)	% Satisfied with Soc & Phy Env. (3)	% Satisfied with Home/School Rel. (4)
Panel A: Parents [N=283]				
TAP	0.424**	2.960	5.386*	5.327**
	(0.211)	(2.494)	(3.062)	(2.457)
Comparison Mean		78.53	73.17	74.10
Panel B: Students [N=293]				
TAP	-0.233	-2.905	-1.725	-2.014
	(0.165)	(1.985)	(2.179)	(1.653)
Comparison Mean		71.91	74.28	81.43
Panel C: Teachers [N=294]				
TAP	0.099	3.361	2.206	-0.283
	(0.204)	(3.813)	(3.031)	(4.106)
Comparison Mean		84.49	88.64	62.11
Controls:				
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes
School Composition	Yes	Yes	Yes	Yes

Notes: This table reports difference-in-differences estimates of the effect of TAP exposure on school climate surveys. The overall satisfaction index, reported in the first column, includes percent of respondents satisfied with (i) learning environment, (ii) social and physical environment, and (iii) home-school relationship. The index is constructed by averaging z-scores of each component. All specifications control for fraction of students who are female, black and free/reduced lunch eligible at the school-by-year level and total school enrollment. Standard errors are clustered at the school level. N represents the sample size.

Table 8: **Benefit-Cost Analysis of TAP**

Panel A: Benefits from Crime Reduction	
Benefits from Reduced Crime	1,952.71
	(2,326.16)
Benefits from Reduced Felony Offenses	1,578.52
	(2242,90)
Panel B: Benefits and Costs of Additional Education	
Benefits from Increased Graduation	2,383.69***
	(707.48)
Cost of Additional Schooling	-346.34***
	(102.79)
Panel C: Net Benefits	
Net Benefits (Reduced Crime+Panel B)	3,990.06
Net Benefits (Reduced Felony Offenses+Panel B)	3,615.87
Total Average Cost of TAP Per Student	250

Notes: This table reports benefits and costs of the TAP program. The social cost estimates of crime come from Miller et al. (1996) and were based on per victim cost values. Benefits associated with earnings are obtained using work-life estimates from Julian and Kominski (2011) and increased life expectancy values reported in Cutler and Lleras-Muney (2006). Finally, the cost of an extra year of school is proxied by expenditures per pupil, which is averaged over 2006-2016. All specifications include the same fixed effects and controls as the main specification. Standard errors are clustered at the school level. ***significant at 1%.

Online Appendix

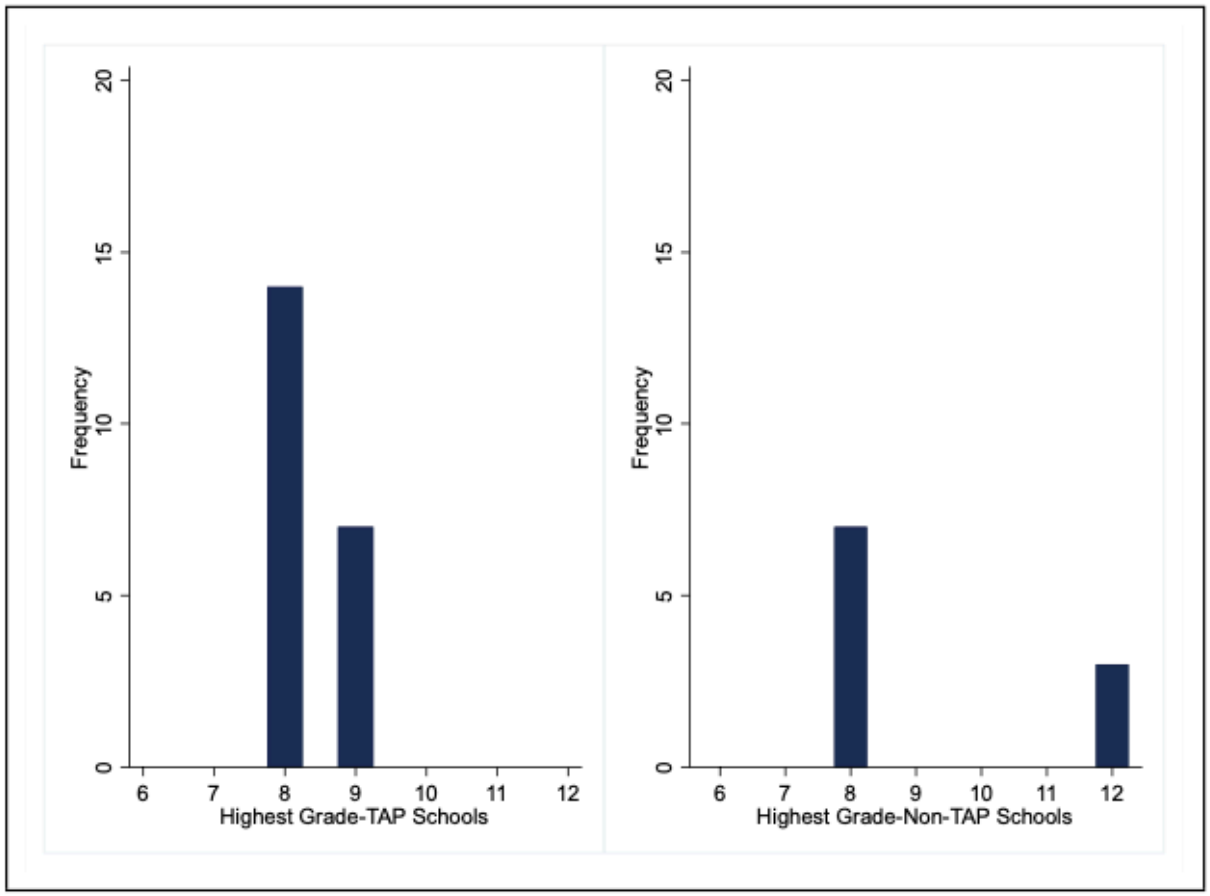
The Long Run Effects of a Comprehensive Teacher Performance Pay Program on Student Outcomes

Sarah Cohodes
Ozkan Eren
Orgul Ozturk

December 2023

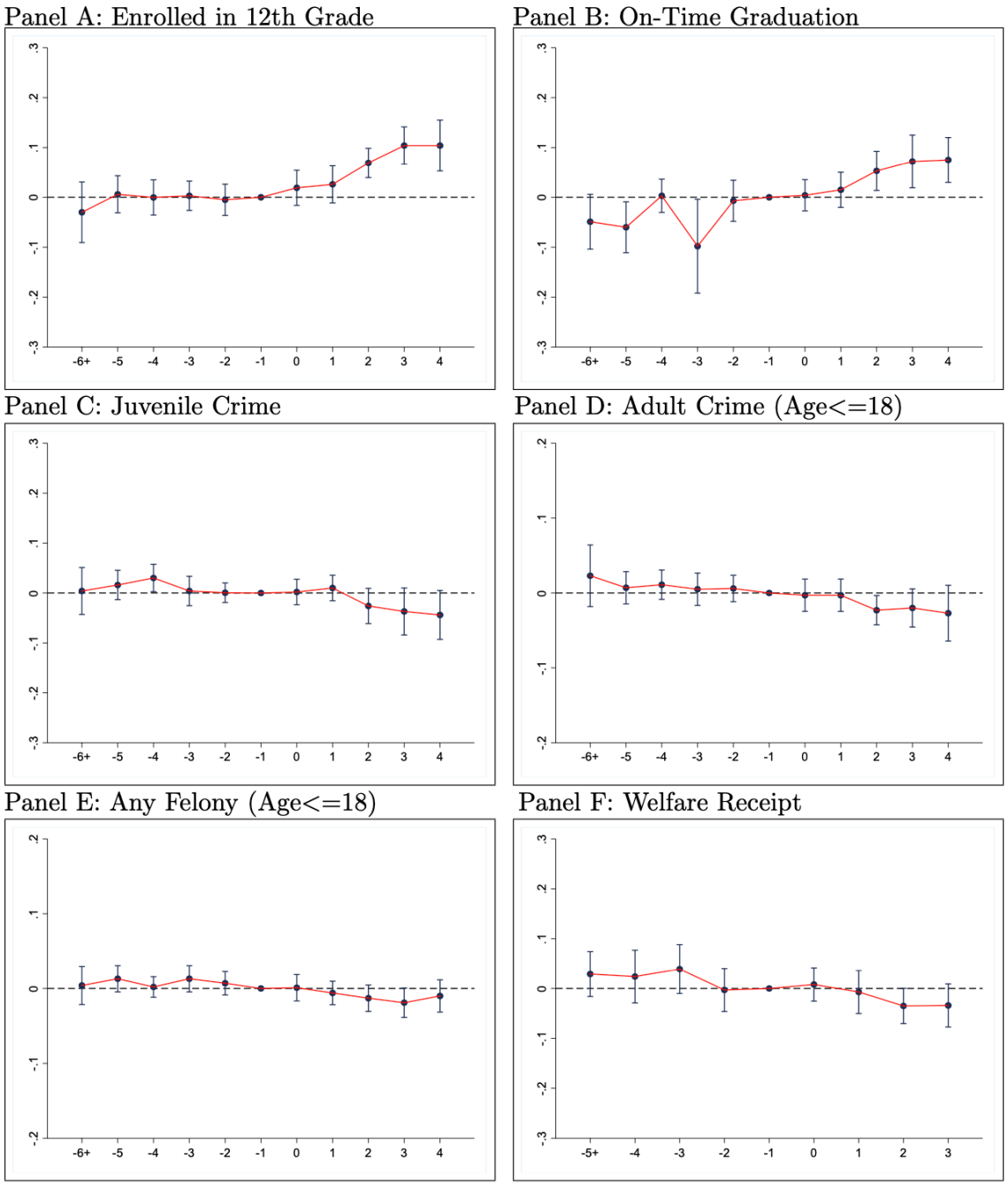
Appendix A: Additional Tables and Figures

Figure A.1: Distribution of Grade Configuration of TAP and Comparison Schools by Highest Grade Offered



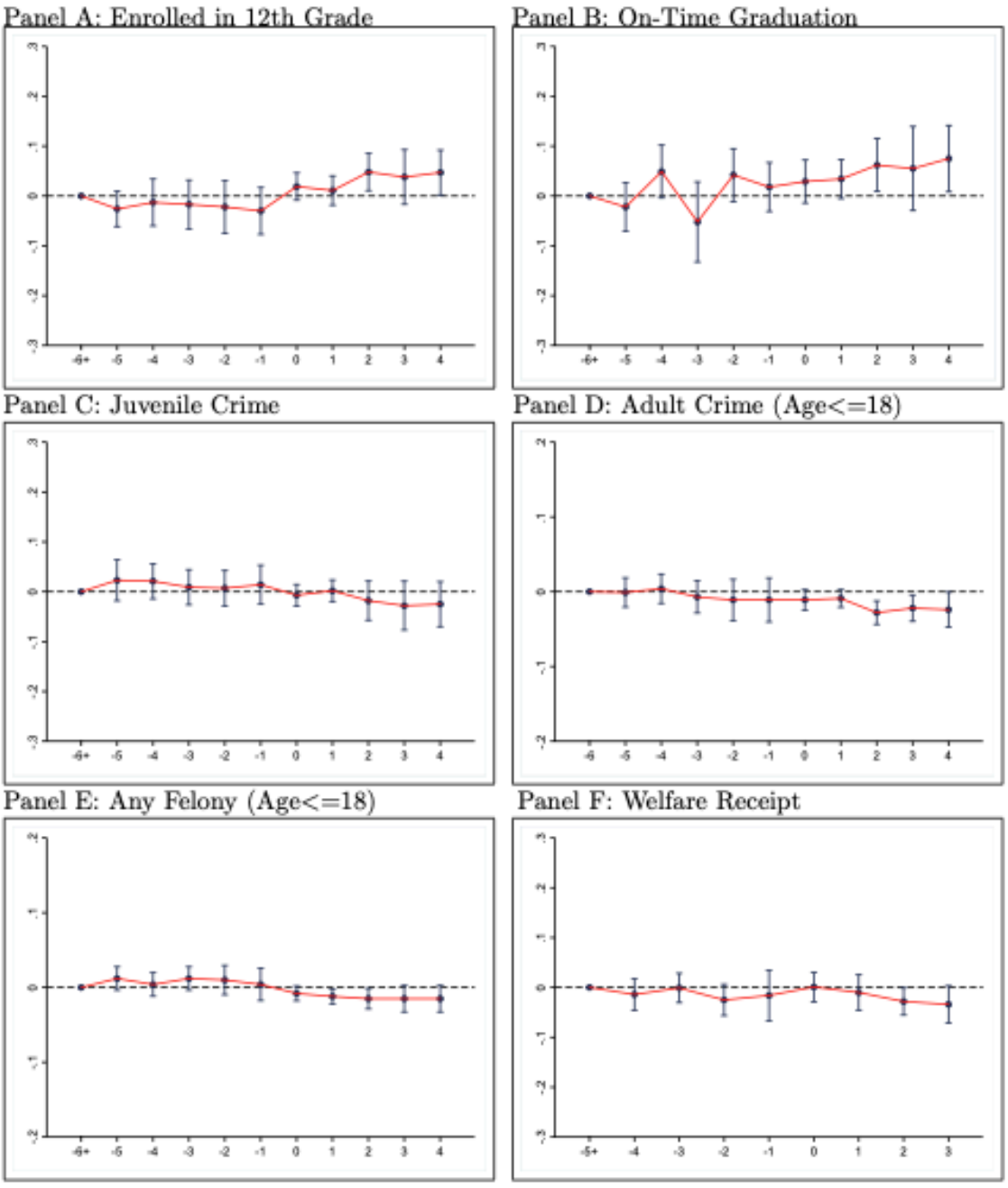
Notes: There are 21 TAP and 10 comparison schools in the analysis sample. The matched comparison group is constructed from the pool of all schools, whose grade configuration includes eighth grade, using baseline school characteristics. The top 5% of schools based on propensity scores comprise the comparison group.

Figure A.2: Event Study Estimates of the Effect of TAP on Long-Run Outcomes: Alternative Estimates



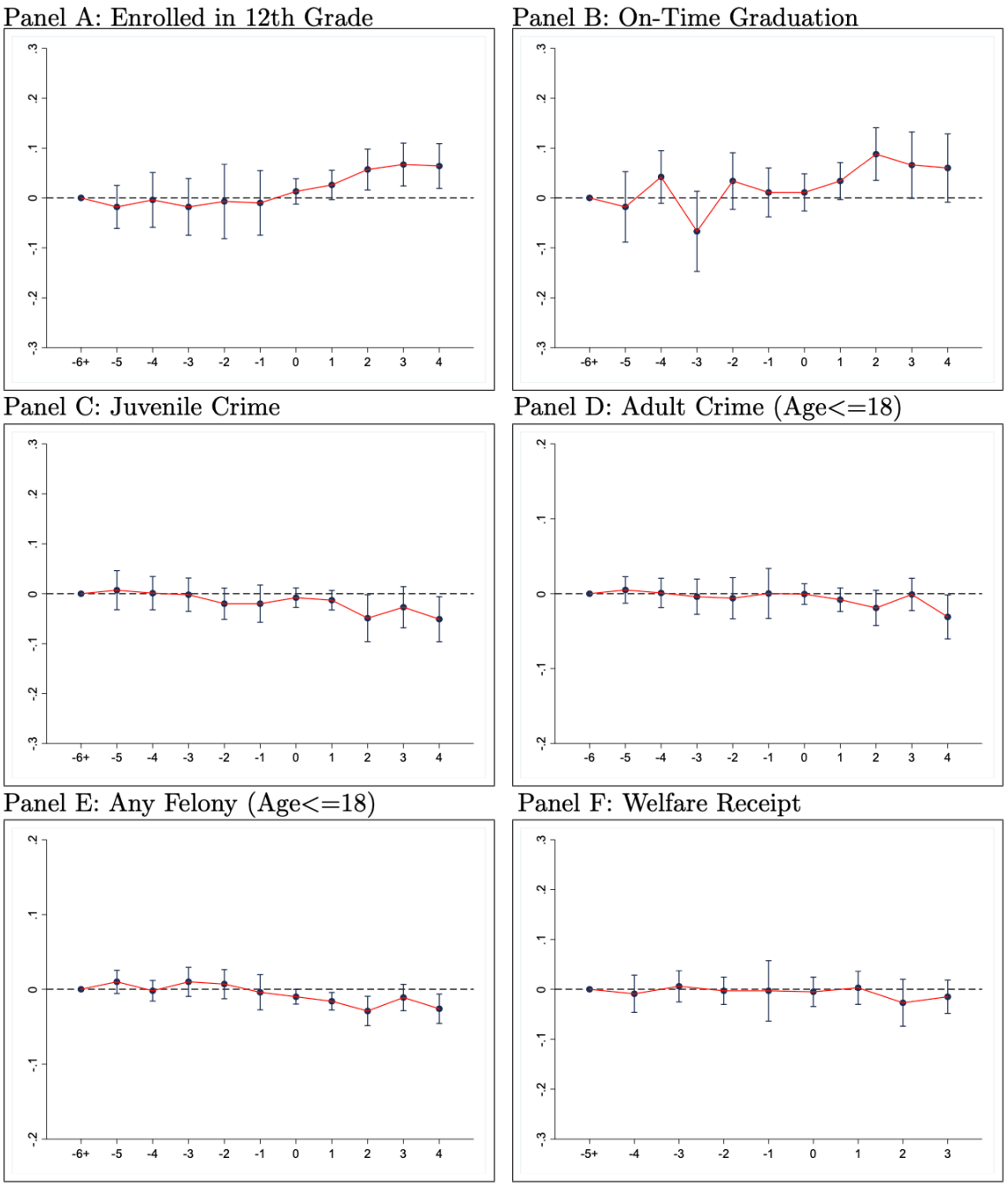
Notes: The coefficient estimates in each panel are obtained using interaction weighted estimator from Sun and Abraham (2021). Each panel shows coefficient estimates and 95% confidence intervals based on standard errors clustered at the school level. Year prior to TAP implementation is the omitted category.

Figure A.3: Event Study Estimates of the Effect of TAP on Long-Run Outcomes: Using the Top 10% of Comparison Schools Based on Their Propensity Scores



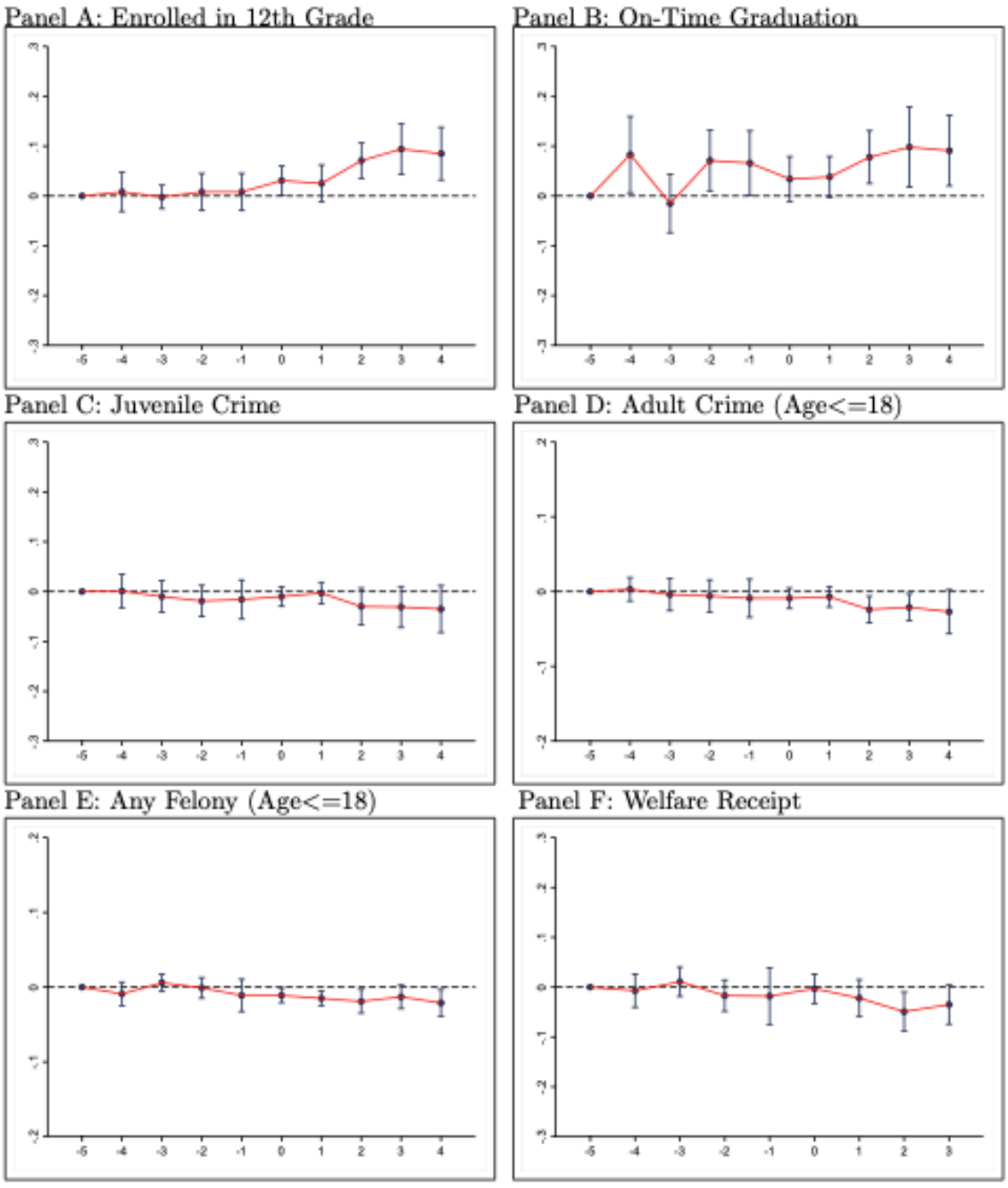
Notes: This figure shows event study estimates for various outcomes obtained using the imputation estimator from Borusyak et al. (2022). The comparison group is defined by selecting the top 10% of comparison schools based on their propensity scores. Each panel shows coefficient estimates and 95% confidence intervals based on standard errors clustered at the school level. Six or more years before TAP adoption ($\tau \leq -6$) is the omitted category.

Figure A.4: Event Study Estimates of the Effect of TAP on Long-Run Outcomes: Future Adopters as Control Group



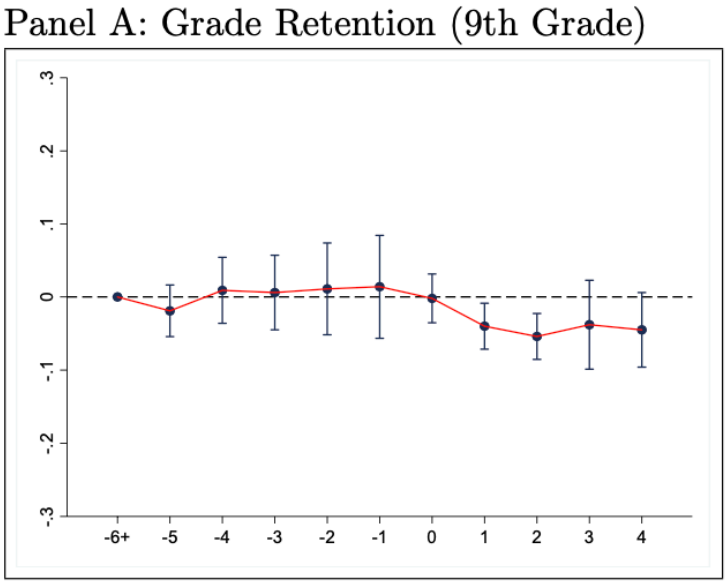
Notes: This figure shows event study estimates for various outcomes obtained using the imputation estimator from Borusyak et al. (2022). The comparison group comprises future adopters, schools adopting TAP outside our study window (2012 or beyond). Each panel shows coefficient estimates and 95% confidence intervals based on standard errors clustered at the school level. Six or more years before TAP adoption ($\tau \leq -6$) is the omitted category.

Figure A.5: Event Study Estimates of the Effect of TAP on Long-Run Outcomes: Balance in Event Time



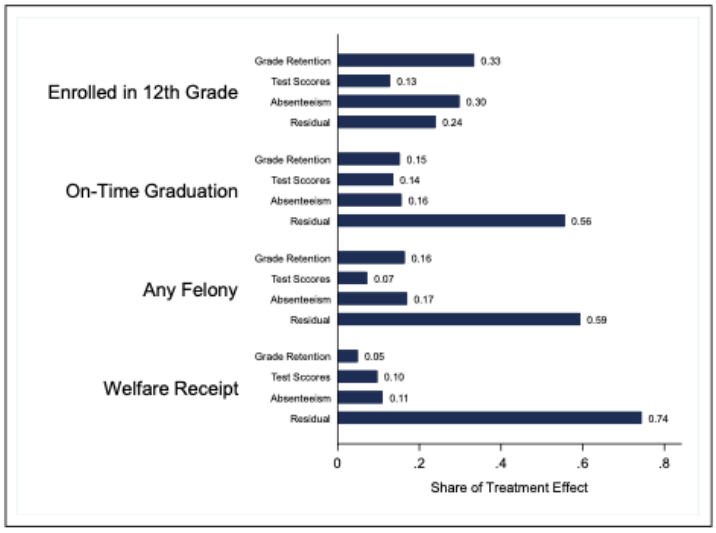
Notes: This figure shows event study estimates for various outcomes obtained using the imputation estimator from Borusyak et al. (2022). Each panel shows coefficient estimates and 95% confidence intervals based on standard errors clustered at the school level. The event window across panels is limited to within 5 years of TAP adoption.

Figure A.6: Event Study Estimates of the Effect of TAP on 9th Grade Retention



Notes: This figure shows event study estimates for 9th grade retention obtained using the imputation estimator from Borusyak et al. (2022). The panel shows coefficient estimates and 95% confidence intervals based on standard errors clustered at the school level. Six or more years before TAP adoption ($\tau \leq -6$) is the omitted category.

Figure A.7: Fraction of TAP Treatment Effect Explained by Intermediate Outcomes



Notes: This figure shows the fraction of TAP treatment effect on longer-run outcomes (educational, criminal justice and economic self-sufficiency) explained by high school outcomes (grade retention, test scores on exit exams and total number of days absent).

Table A.1: **Timeline of TAP Implementation, Exposure, and Data Availability**

	Cohort	Academic Outcomes (Normal Grade Progression Shown)				Criminal Justice Outcomes	Self-Sufficiency Outcomes
	8th Grade	9th Grade	10th Grade	11th Grade	12th Grade	By Birth Year	By Birth Year
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: TIF 1 (2007)							
1st Post-Adoption Cohort	2007-08 1 exposure year	2008-09	2009-10	2010-11	2011-12	Up to Age 18	Ages 18-22
2nd Post-Adoption Cohort	2008-09 2 exposure years	2009-10	2010-11	2011-12	2012-13	Up to Age 18	Ages 18-22
3rd Post-Adoption Cohort	2009-10 3 exposure years	2010-11	2011-12	2012-13	2013-14	Up to Age 18	Ages 18-22
4th Post-Adoption Cohort	2010-11 3 exposure years	2011-12	2012-13	2013-14	2014-15	Up to Age 18	Ages 18-22
5th Post-Adoption Cohort	2011-12 3 exposure years	2012-13	2013-14	2014-15	2015-16	Up to Age 18	
Panel B: TIF 2 (2008)							
1st Post-Adoption Cohort	2008-09 1 exposure year	2009-10	2010-11	2011-12	2012-13	Up to Age 18	Ages 18-22
2nd Post-Adoption Cohort	2009-10 2 exposure years	2010-11	2011-12	2012-13	2013-14	Up to Age 18	Ages 18-22
3rd Post-Adoption Cohort	2010-11 3 exposure years	2011-12	2012-13	2013-14	2014-15	Up to Age 18	Ages 18-22
4th Post-Adoption Cohort	2011-12 3 exposure years	2012-13	2013-14	2014-15	2015-16	Up to Age 18	
5th Post-Adoption Cohort	2012-13 3 exposure years	2013-14	2014-15	2015-16	2016-17	Up to Age 18	
Panel C: TIF 3(2010)							
1st Post-Adoption Cohort	2010-11 1 exposure year	2011-12	2012-13	2013-14	2014-15	Up to Age 18	
2nd Post-Adoption Cohort	2011-12 2 exposure years	2012-13	2013-14	2014-15	2015-16	Up to Age 18	
3rd Post-Adoption Cohort	2012-13 3 exposure years	2013-14	2014-15	2015-16	2016-17	Up to Age 18	

Notes: This table reports the staggered TAP implementation used as the source of variation in the difference-in-difference estimation strategy and data availability associated with each TAP cohort. Exposure years refers to the number of years an 8th grade cohort may have been exposed to the TAP intervention in middle school, assuming a 6-8 grade configuration and that a student was in the school for all of middle school.

Table A.2: **Summary Statistics by School Type, Baseline (2002-2003) Academic Year**

	TAP Schools	Matched Comparison Schools	All Other Schools	Future Adopters	Test of Equality Cols. (1) vs. (2) (p-value)	Test of Equality Cols. (1) vs. (3) (p-value)	Test of Equality Cols. (1) vs. (4) (p-value)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Share Female Students (8th Grade)	0.495	0.491	0.498	0.482	0.82	0.81	0.54
Share Black Students (8th Grade)	0.588	0.740	0.403	0.584	0.11	0.00	0.98
Share White Students (8th Grade)	0.386	0.243	0.561	0.376	0.13	0.00	0.94
Share Free/Reduced Lunch Students (8th Grade)	0.693	0.829	0.499	0.664	0.04	0.00	0.74
Total School Enrollment	516.85	455.40	665.57	515.17	0.46	0.02	0.99
Student Attendance Rate (%)	94.59	94.46	95.16	94.83	0.82	0.12	0.72
% of Students Suspended/Expelled	4.83	3.19	2.69	1.05	0.50	0.02	0.23
# of Teachers in the School	35.90	33.30	44.27	34.83	0.62	0.03	0.87
% of Teachers with an Advanced Degree	46.93	50.78	46.38	45.17	0.46	0.83	0.79
% of Continuing Contract Teachers	80.45	82.44	79.52	78.48	0.71	0.73	0.76
% of Teachers Satisfied w/ Environment	83.18	85.30	84.55	85.57	0.71	0.71	0.74
Number of Schools	21	10	217	6			

Notes: This table reports school characteristics for TAP schools and comparison schools. The matched comparison sample is constructed from the pool of all schools, whose grade configuration includes eighth grade, using baseline school characteristics listed above. The top 5% of schools based on propensity scores comprise the comparison group (Column 2). Column 3 presents all other schools in the state whose grade configuration include eighth grade. Column 4 includes future TAP adopters which are schools adopting TAP in 2012 (or beyond) as part of TIF 4.

Table A.3: **Predicting TAP Adoption Year**

	TAP Adoption	
	Early Adopters	Late Adopters
	(1)	(2)
Fraction of Female Students (8th Grade)	-1.655	2.602
	(2.589)	(1.822)
Fraction of Black Students (8th Grade)	-0.145	-0.235
	(0.858)	(0.604)
Fraction of Free/Reduced Lunch Students (8th Grade)	-0.053	-1.134
	(1.323)	(0.932)
Total School Enrollment	0.002	-0.001
	(0.002)	(0.001)
Student Attendance Rate (%)	-0.035	0.041
	(0.098)	(0.069)
Percent of Students Suspended/Expelled	0.015	-0.003
	(0.022)	(0.015)
Total Number of Teachers in the School	-0.019	0.003
	(0.032)	(0.022)
Percent of Teachers with an Advanced Degree	0.008	-0.006
	(0.009)	(0.006)
Percent of Continuing Contract Teachers	-0.005	0.000
	(0.012)	(0.008)
Percent of Teachers Satisfied with Social and Physical Environment	-0.003	-0.003
	(0.011)	(0.008)
Baseline (5th Grade) Composite Score	0.473	-0.687
	(1.048)	(0.738)
F-test (p-value)	0.91	0.46
Sample Size	31	31

Notes: This table tests whether the timing of TAP adoption is associated with school characteristics from the baseline (2002-2003) academic year. The early adopters are schools implementing TAP in the 2007-2008 and 2008-2009 academic years, while the late adopters implemented TAP in the 2010-2011 academic year. The F-test p-value comes from a test that the coefficients shown are jointly equal to zero.

Table A.4: **Effect of TAP on Long-Run Outcomes-Treatment Defined at the District Level Based on TAP Adoption**

	Enrolled in 12th Grade (1)	Graduated from HS in 4 Years (2)	Any Felony (Age≤18) (3)	Welfare Receipt (4)
Panel A: All Districts				
TAP	-0.009	0.007	-0.006	0.005
	(0.014)	(0.018)	(0.004)	(0.011)
Comparison Mean	0.698	0.679	0.043	0.576
Sample Size	550,826	500,781	550,826	435,680
Panel B: Comparison Districts				
TAP	0.006	0.016	-0.007	-0.001
	(0.014)	(0.018)	(0.004)	(0.011)
Comparison Mean	0.678	0.664	0.056	0.624
Sample Size	114,342	103,335	114,342	81,315
Controls:				
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes

Notes: This table replicates the estimates from Tables 4 but with TAP exposure defined at the district level rather than the school level. The treatment indicator defined at the district level takes the value one if any of the schools in the district adopted TAP by that particular year. Panel A includes all school districts, while Panel B limits the control group to districts associated with matched control schools. Standard errors are clustered at the school level. See Table 4 and the text for further details.

Table A.5: **Effect of TAP on Juvenile and Adult Felony and Non-Felony Crime**

	Juv. Felony (1)	Juv. Non-Felony (2)	Adult Felony (3)	Adult Non-Felony (4)	Any Non-Felony (Age≤18) (5)
Panel A: Differences-in-Differences Estimates					
TAP	-0.008**	-0.009	-0.006*	-0.011	-0.014
	(0.004)	(0.011)	(0.003)	(0.007)	(0.012)
Panel B: Semi-Dynamic Model Estimates					
1st Postadoption Cohort	-0.006	-0.009	-0.004	-0.010	-0.012
	(0.005)	(0.009)	(0.004)	(0.007)	(0.011)
2nd Postadoption Cohort	-0.009**	0.000	-0.006*	-0.004	-0.005
	(0.004)	(0.012)	(0.003)	(0.008)	(0.012)
3rd Postadoption Cohort	-0.009*	-0.028*	-0.011**	-0.024**	-0.036*
	(0.005)	(0.017)	(0.005)	(0.009)	(0.019)
4th Postadoption Cohort	-0.009	-0.022	-0.004	-0.025***	-0.030
	(0.008)	(0.021)	(0.006)	(0.009)	(0.020)
5th Postadoption Cohort	-0.014**	-0.033	-0.010*	-0.025*	-0.048*
	(0.007)	(0.023)	(0.005)	(0.015)	(0.025)
Comparison Mean	0.029	0.130	0.020	0.060	0.164
Sample Size	43,062	43,062	43,062	43,062	43,062
Controls:					
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes	Yes

Notes: This table reports difference-in-differences and semi-dynamic estimates of the effect of TAP exposure on disaggregated criminal activity outcomes. The coefficient estimates in Panel B are obtained using imputation estimator from Borusyak et al. (2022). Standard errors are clustered at the school level. The dependent variable takes the value one if student was ever arrested for a juvenile/adult felony or non-felony crime. See notes to Table 4 and the text for further details. * significant at 10%, ** significant at 5%.

Table A.6: **Effect of TAP on Additional Outcomes**

		Juv/Adult Crime by Type				Composite Test	Long-Run
	Attrition	Violent	Alc-Drug	Property	Other	Score (5th Grade)	Outcome Index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
TAP	0.0004	-0.009	-0.009**	-0.009	-0.014	-0.027	0.047***
	(0.0041)	(0.007)	(0.003)	(0.007)	(0.010)	(0.028)	(0.016)
Comparison Mean	0.045	0.056	0.031	0.059	0.115	-0.488	0.004
Sample Size	43,092	43,062	43,062	43,062	43,062	33,459	43,062
Controls:							
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports difference-in-differences estimates of the effect of TAP exposure on various outcome variables. Standard errors are clustered at the school level. Attrition indicator takes the value one if student had not ever enrolled in ninth grade in a South Carolina public school. The long-run outcome index includes binary indicators for enrollment in twelfth grade, on-time high school graduation, being arrested of a felony offense by age 18 or earlier and enrollment in social programs as an adult between ages 18 and 22. The index is constructed by averaging z-scores of each component (the binary indicators for ever being arrested of a felony offense and enrollment in social programs are reverse coded). See notes to Table 4 and the text for further details. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table A.7: **Effect of TAP on Long-Run Outcomes: Interaction Weighted Estimator (Sun and Abraham 2021)**

	Enrolled in 12th Grade (1)	Graduated from HS in 4 Years (2)	Any Juvenile Crime (3)	Any Adult Crime (4)	Any Felony (Age≤18) (5)	Welfare Receipt (6)
Semi-Dynamic Model Estimates						
1st Postadoption Cohort	0.018	0.026	-0.008	-0.007	-0.010*	-0.007
	(0.013)	(0.021)	(0.010)	(0.008)	(0.006)	(0.015)
2nd Postadoption Cohort	0.025*	0.038**	-0.003	-0.006	-0.015**	-0.025
	(0.013)	(0.018)	(0.011)	(0.008)	(0.006)	(0.018)
3rd Postadoption Cohort	0.059***	0.066***	-0.029	-0.022**	-0.018**	-0.053***
	(0.015)	(0.023)	(0.019)	(0.010)	(0.008)	(0.018)
4th Postadoption Cohort	0.071***	0.078**	-0.027	-0.017	-0.014	-0.041*
	(0.022)	(0.039)	(0.022)	(0.011)	(0.010)	(0.021)
5th Postadoption Cohort	0.071***	0.081***	-0.033	-0.025	-0.022**	-
	(0.023)	(0.031)	(0.024)	(0.016)	(0.010)	
Comparison Mean	0.673	0.664	0.139	0.068	0.045	0.607
Sample Size	43,062	38,253	43,062	43,062	43,062	30,081
Controls:						
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports coefficient estimates obtained using the interaction weighted estimator from Sun and Abraham (2021) which parallel the estimates in Table 4. Standard errors are clustered at the school level. See notes to Table 4 and the text for further details. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table A.8: **Robustness Checks: Effect of TAP on Long-Run Outcomes**

	Enrolled in 12th Grade (1)	Graduated from HS in 4 Years (2)	Any Felony (Age≤18) (3)	Welfare Receipt (4)
Panel A: Exclude Late Adopters				
TAP	0.037**	0.067**	-0.017**	-0.027***
(std)	(0.017)	(0.027)	(0.007)	(0.013)
[sample size]	[33,858]	[29,997]	[33,858]	[30,081]
Panel B: Exclude Early Adopters (TIF 1 Schools)				
TAP	0.049***	0.021*	-0.006	-0.029**
	(0.010)	(0.011)	(0.007)	(0.013)
	[25,495]	[22,753]	[25,495]	[14,118]
Panel C: Control 5th Grade Comp. Test Scores				
TAP	0.045***	0.048**	-0.013**	-0.035***
	(0.012)	(0.020)	(0.006)	(0.010)
	[33,459]	[33,370]	[33,459]	[22,587]
Panel D: Control for Trends by School Characteristics				
TAP	0.029**	0.031*	-0.010**	-0.024*
	(0.012)	(0.017)	(0.004)	(0.012)
	[43,062]	[38,253]	[43,062]	[30,081]
Panel E: Control for District-Specific Linear Pretrends				
TAP	0.058**	0.038	-0.014	-0.026
	(0.023)	(0.032)	(0.009)	(0.020)
	[43,062]	[38,253]	[43,062]	[30,081]
Panel F: Exclude Student and Grade Controls				
TAP	0.031***	0.042**	-0.012**	-0.017
	(0.011)	(0.019)	(0.006)	(0.015)
	[43,062]	[38,253]	[43,062]	[30,081]
Panel G: Exclude Students Enrolled in TAP High Schools				
TAP	0.036**	0.064**	-0.020***	-0.022
	(0.013)	(0.028)	(0.007)	(0.015)
	[28,946]	[25,704]	[28,946]	[21,419]

Robustness Checks: Effect of TAP on Long-Run Outcomes (Continued from Previous Page)

	Enrolled in 12th Grade (1)	Graduated from HS in 4 Years (2)	Any Felony (Age≤18) (3)	Welfare Receipt (4)
Panel H: Control for GEAR UP				
TAP	0.038***	0.042**	-0.015***	-0.027***
	(0.010)	(0.018)	(0.005)	(0.013)
	[43,062]	[38,253]	[43,062]	[30,081]
Panel I: Comparison Group-Propensity Score Top 10%				
TAP	0.024**	0.035*	-0.011**	-0.017
	(0.010)	(0.020)	(0.005)	(0.012)
	[56,335]	[49,902]	[56,335]	[41,944]
Panel J: Comparison Group-Future Adopters				
TAP	0.028**	0.032*	-0.017***	-0.013
	(0.012)	(0.018)	(0.005)	(0.013)
	[38,367]	[34,017]	[38,367]	[27,382]
Panel K: Treatment Intensity				
TAP	0.059***	0.070***	-0.027***	-0.064***
	(0.020)	(0.020)	(0.009)	(0.021)
	[43,062]	[38,253]	[43,062]	[30,081]
Panel L: Clustering at the School-by-Cohort Level				
TAP	0.035***	0.038**	-0.014**	-0.027**
	(0.009)	(0.016)	(0.004)	(0.011)
	[43,062]	[38,253]	[43,062]	[30,081]
Panel M: Wild Bootstrap				
TAP	0.035**	0.038**	-0.014**	-0.027*
(p-value)	(0.016)	(0.041)	(0.022)	(0.065)
	[43,062]	[38,253]	[43,062]	[30,081]
Controls:				
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes

Notes: This table reports difference-in-differences estimates of the effect of TAP exposure on key outcome variables from several robustness tests. Panel A excludes schools adopting TAP in the 2010-2011 academic year; Panel B excludes schools adopting TAP in the 2007-2008 academic year. Panel C controls for 5th grade baseline composite test scores. Panel D adds interaction of baseline school level controls with a linear trend. Panel E controls for district-specific linear pretrends. Panel F excludes student and grade level controls. Panel G excludes all students ever enrolling in a TAP high school. Panel H conditions on schools' GEAR UP status. Panel I expands the matched control group and selects the top 10% of schools based on propensity scores. Panel J uses future TAP adopters (post-2012) as an alternate control group. Panel K estimates the models continuously using potential years of exposure to TAP. Standard errors are clustered at the school level in Panels A-K, while they are clustered at the school-by-cohort level in Panel L. Panel M reports p-values associated with test of significance for each coefficient estimate using the wild bootstrap procedure clustered at the school level. Sample sizes are reported in square brackets. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table A.9: **Effect of TAP on Long-Run Outcomes-Using Different Grade Spans**

	Enrolled in 12th Grade (1)	Graduated from HS in 4 Years (2)	Any Felony (Age$\leq$ 18) (3)	Welfare Receipt (4)
Panel A: Exclude Schools with Grade Config >9th Grade				
TAP	0.033***	0.036**	-0.013**	-0.028**
	(0.011)	(0.018)	(0.005)	(0.013)
Comparison Mean	0.669	0.660	0.047	0.617
Sample Size	41,390	36,784	41,390	28,681
Panel B: Exclude Schools with Grade Config > 9th Grade or Grade Config < 5th Grade				
TAP	0.039***	0.047**	-0.010**	-0.021
	(0.010)	(0.021)	(0.005)	(0.014)
Comparison Mean	0.669	0.660	0.047	0.617
Sample Size	36,028	31,992	36,028	23,869
Controls:				
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes
School Fixed Effects (9th Grade)	Yes	Yes	Yes	Yes

Notes: This table replicates the key estimates from Table 4 by excluding schools whose highest grade configuration spans grades above ninth grade (Panel A). Panel B additionally excludes schools whose lowest grade configuration spans below fifth grade. Standard errors are clustered at the school level. See Table 4 and the text for further details. ** significant at 5%, *** significant at 1%.

Table A.10: **Placebo Effect of TAP on Long-Run Outcomes-Using Pre-TAP Adoption Years**

	Enrolled in 12th Grade (1)	Any Felony (Age$\leq$ 18) (2)	Welfare Receipt (3)
TAP	-0.006	0.007	-0.001
	(0.014)	(0.005)	(0.012)
Comparison Mean	0.652	0.017	0.634
Sample Size	34,701	34,701	27,523
Controls:			
Cohort and School Fixed Effects	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes

Notes: This table reports a placebo effect of TAP with TAP adoption indicated four years earlier than actual exposure. The analysis sample is restricted to students enrolled in eighth grade for the first time between the 2000-2001 and 2007-2008 academic years. The models are estimated as if treated schools first adopted TAP in 2003 (rather than 2007), with schools adopting TAP t years after 2007 as if they adopted in 2003+ t and use post-adoption data from the first 4 cohorts of each placebo TAP wave (e.g., 8th grade cohorts from 2003 to 2006 for schools adopting TAP in 2003). Standard errors are clustered at the school level. See notes to Table 4 and the text for further details.

Table A.11: **Effect of TAP on Long-Run Outcomes, by Subgroups**

	Female Students	Male Students	Black Students	White Students
Panel A: Enrolled in 12th Grade				
TAP	0.036**	0.034**	0.041**	0.021
	(0.013)	(0.015)	(0.015)	(0.017)
Comparison Mean	0.740	0.601	0.690	0.634
Test of Equality by Subgroups (p-value)	0.94		0.38	
Sample Size	21,160	21,902	25,410	15,976
Panel B: Graduated from HS in 4 Years				
TAP	0.044**	0.032	0.053**	0.022
	(0.017)	(0.023)	(0.022)	(0.023)
Comparison Mean	0.733	0.596	0.679	0.628
Test of Equality by Subgroups (p-value)	0.69		0.32	
Sample Size	18,785	19,468	22,506	14,181
Panel C: Any Felony (Age$\leq$18)				
TAP	-0.004	-0.023**	-0.019**	-0.009
	(0.004)	(0.009)	(0.008)	(0.005)
Comparison Mean	0.012	0.078	0.049	0.037
Test of Equality by Subgroups (p-value)	0.06		0.34	
Sample Size	21,160	21,902	25,410	15,976
Panel D: Welfare Receipt				
TAP	-0.022*	-0.031	-0.035**	-0.011
	(0.013)	(0.025)	(0.016)	(0.024)
Comparison Mean	0.663	0.552	0.673	0.431
Test of Equality by Subgroups (p-value)	0.74		0.41	
Sample Size	14,833	15,248	19,396	9,440
Controls:				
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes

Notes: This table reports difference-in-differences estimates of the effect of TAP exposure for different student subgroups. The test p-values come from tests of equality of the coefficient estimates between subgroups. Standard errors are clustered at the school level. See Table 4 and the text for further details.** significant at 5%.

Table A.12: **Mediation Analysis-TAP, Educational Attainment and Felony Crime**

	Enrolled in 12th Grade	Enrolled in 12th Grade	Graduated from HS in 4 Years	Graduated from HS in 4 Years	Adult Felony Crime	Adult Felony Crime	Welfare Receipt	Welfare Receipt
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TAP	0.044***	0.010	0.078*	0.044	-0.007**	-0.004	-0.054***	-0.022
	(0.013)	(0.012)	(0.041)	(0.043)	(0.003)	(0.003)	(0.018)	(0.016)
Grade Retention		-0.359***		-0.290***		0.028***		0.065***
		(0.016)		(0.022)		(0.006)		(0.015)
Composite Test Score		0.078***		0.147***		-0.007***		-0.073***
		(0.004)		(0.008)		(0.001)		(0.005)
Absenteeism		-0.0044***		-0.0041***		0.0004***		0.002***
		(0.0003)		(0.0004)		(0.0001)		(0.0003)
Control Mean	0.775		0.767					
Sample Size	19,224	19,224	19,224	19,224	19,224	19,224	12,779	12,779
Controls:								
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Odd numbered columns reproduce the baseline estimates using student observations with non-missing data on intermediate outcomes while even numbered columns report the estimated effects of TAP using the modified version of the DiD specification. Standard errors are clustered at the school level. See notes to Tables 4 and 5 and the text for further details.

Table A.13: **Effect of TAP on Long-Run Outcomes: Excluding TAP Schools Experiencing High Average Teacher Turnover**

	% of Teachers Returning School from Previous Year	Enrolled in 12th Grade	Graduated from HS in 4 Years	Any Felony (Age¡=18)	Welfare Receipt
	(1)	(2)	(3)	(4)	(5)
TAP	0.005	0.045***	0.049*	-0.015**	-0.036**
	(1.197)	(0.011)	(0.026)	(0.006)	(0.013)
Comparison Mean	82.21	0.673	0.664	0.045	0.607
Sample Size	224	33,833	30,117	33,833	24,292
Controls:					
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes	Yes
School Composition	Yes	No	No	No	No
Student Characteristics	No	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	No	Yes	Yes	Yes	Yes

Notes: The analysis sample excludes TAP schools which experience an average change in teacher turnover rates by more than 5 percentage points (in absolute value) between pre- and post-adoption periods. Standard errors are clustered at the school level. See Table 4 and the text for further details. * significant at 10% ** significant at 5%, *** significant at 1%.

Table A.14: **Effect of TAP on Long-Run Outcomes: Controlling for Ninth-Grade School Fixed Effects**

	Enrolled in 12th Grade	Graduated from HS in 4 Years	Any Felony (Age≤18)	Welfare Receipt
	(1)	(2)	(3)	(4)
TAP	0.033***	0.035**	-0.012**	-0.024*
	(0.010)	(0.017)	(0.005)	(0.013)
Comparison Mean	0.700	0.689	0.044	0.616
Sample Size	40,774	36,222	40,774	28,343
Controls:				
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes	Yes
School Fixed Effects (9th Grade)	Yes	Yes	Yes	Yes

Notes: This table replicates the estimates from Table 4 but with ninth grade school fixed effects. Standard errors are clustered at the school level. All specifications include the same fixed effects and controls as the main specification. See Table 4 and the text for further details. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table A.15: **Effect of TAP on Leadership Change and Controlling for School Principal's Years at the School**

	Leadership Change (1)	Enrolled in 12th Grade (2)	Graduated from HS in 4 Years (3)	Any Felony (Age<=18) (4)	Welfare Receipt (5)
TAP	-0.000	0.031***	0.040*	-0.014***	-0.029**
	(0.083)	(0.010)	(0.021)	(0.005)	(0.012)
Comparison Mean	24.74	0.673	0.664	0.045	0.607
Sample Size	302	43,062	38,253	43,062	30,081
Controls:					
Cohort and School Fixed Effects	Yes	Yes	Yes	Yes	Yes
School Composition	Yes	No	No	No	No
Student Characteristics	No	Yes	Yes	Yes	Yes
Grade Composition (8th Grade)	No	Yes	Yes	Yes	Yes

Notes: This table reports difference-in-differences and semi-dynamic estimates of the effect of TAP exposure on leadership change and key outcomes, controlling for leadership change. Standard errors are clustered at the school level. The dependent variable in Column (1) takes the value one if school's principal changed from (t-1) to (t). The specifications in Columns (2)-(5) control for school principal's years at the school. See notes to Table 4 and the text for further details. * significant at 10%, ** significant at 5%.

Appendix B: An Example of the TAP Compensation System

Consider a middle school with 12 teachers. Assume that the school allocates $2,000 per teacher to establish the TAP award fund ($24,000 in total). We first concentrate on the classroom observations component of the performance pay. Scores are measured on a five-point scale based on the TAP teacher observation rubric. These observation rubrics captured a range of practices in domains such as the teacher's planning and preparation for learning, delivery of instruction, classroom management and environment.

Table B.1 presents a scenario where eight teachers earned classroom observation scores less than 2.5. In our simple example, a teacher must earn at least a score of 2.5 in order to be eligible for teaching practices domain of the pay scheme. The scale of the classroom observation scores and the total number of teachers at each scale are reported in Columns 1 and 2, respectively. Column 3 displays the fixed pay ratios associated with each score. Pay ratios are used to weight scores so that teachers with higher scores earn larger shares from the award pool, e.g., a teacher with a score of 5.0 earns seven times more than a teacher earning a score of 2.5. Multiplying Columns 2 and 3 and dividing the total award fund designated for teaching practices category (40 percent of the fund-$9,600) by the sum from this product (Column 4) yields the award amount per teacher at a pay ratio of 1 which is equal to $640.

Next, the total award fund that is designated for classroom achievement growth is $7,200 (30 percent of the fund). Table B.2 depicts a scenario for teacher value-added scores. Columns 1 and 2 are as previously defined and Column 3 represents the pay ratios for teacher-value added scores. For ease of interpretation, value-added scores are also converted to a five-point scale. A score of 4 (2) represents more (less) than one year of full year academic growth. Following the same metric, we can calculate the award amount per teacher at a pay ratio of 1 as $424.

Finally, the remaining 30 percent of the total bonus award allocation depends on the school-level value-added scores ($7,200). Table B.3 presents the distribution of award eligibility for school-level achievement growth. A school achieving level 5 performance would receive all the award money allocated, while a school with level 4 performance would receive 75 percent of the money, and all other school-level achievement growth scales are defined similarly.

Suppose there is a teacher with an average classroom observation score of 3.5, a value-added score of 3 and teaches in a TAP school in which the school-level achievement gain is 4.0. This teacher receives a total performance pay of $2,794 (1,920+424+450).

Table B.1: Classroom Observation Scores and Pay Ratios

Observation Scores (1)	Number of Teachers (2)	Pay Ratios (3)	Number of Teachers × Pay Ratios (4)
1	4	0	0
2	4	0	0
2.5	0	1	0
3	1	2	2
3.5	2	3	6
4	0	5	0
4.5	0	6	0
5	1	7	7
Total	12		15

Notes: This table shows observation scores and pay ratios for a hypothetical school with 12 teachers. Pay ratios are fixed.

Table B.2: Teacher Value-Added Scores and Pay Ratios

Teacher Value-Added Scores (Test Gains) (1)	Number of Teachers (2)	Pay Ratios (3)	Number of Teachers × Pay Ratios (4)
1	0	0	0
2	9	0	0
3	1	1	1
4	1	6	6
5	1	10	10
Total	12		17

Notes: This table shows teacher value-added scores and pay ratios for a hypothetical school with 12 teachers. Pay ratios are fixed.

Table B.3: School-Level Value-Added Scores and Pay Ratios

School Value-Added Scores (Test Gains) (1)	Percent of Award Allocated for School (2)
1	0
2	0
3	50%
4	75%
5	100%

Notes: This table shows school value-added scores and pay ratios for a hypothetical school with 12 teachers. Pay ratios are fixed.