

NBER WORKING PAPER SERIES

TEACHER PERFORMANCE PAY, COACHING, AND LONG-RUN STUDENT OUTCOMES

Sarah Cohodes
Ozkan Eren
Orgul Ozturk

Working Paper 31056
<http://www.nber.org/papers/w31056>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2023

We are grateful to Richard Mansfield, Katherine Parham Malhotra, Jonah Rockoff, Eric Taylor, Yotam Shem-Tov and seminar participants at Vanderbilt University, the University of South Carolina 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2023 by Sarah Cohodes, Ozkan Eren, and Orgul Ozturk. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Teacher Performance Pay, Coaching, and Long-Run Student Outcomes
Sarah Cohodes, Ozkan Eren, and Orgul Ozturk
NBER Working Paper No. 31056
March 2023
JEL No. H75,I21,J32,J45

ABSTRACT

This paper examines the effects of a comprehensive performance pay program for teachers 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 identify causal effects of the school reform. The program 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, and that program benefits far exceeded its costs. We propose mechanisms for observed long-run effects and provide evidence consistent with these explanations. Several robustness checks and placebo tests support our findings.

Sarah Cohodes
Teachers College
Columbia University
525 West 120th Street
New York, NY 10027
and NBER
cohodes@tc.columbia.edu

Orgul Ozturk
1014 Greene Street
Columbia SC
29208
odozturk@moore.sc.edu

Ozkan Eren
Department of Economics
University of California, Riverside
Sproul Hall
Riverside, CA 92521
ozkan.eren@ucr.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 necessarily increases 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 examining the effectiveness of teacher incentives, it is further important to gauge whether short-term effects (if any) persist or fade out and whether these effects translate into meaningful change in long-run outcomes.

In this paper, we study the medium- and long-run effects of performance-based compensation 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, 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 South Carolina's 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. For teachers in grades and subjects in which state assessments are not administered, bonus allocation is based on school achievement growth and teacher practices (described in more detail below). In this regard, South Carolina 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

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

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)).

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 pre-trends as well as endogenous mobility, conditioning on district-specific trends and experimenting with alternate comparison samples) and different placebo tests supporting our identifying assumption.

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.

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. Further examination provides suggestive evidence of TAP schools attracting lower quality and less experienced teachers relative to school leavers, though this provided schools an opportunity to bring in teachers who may be responsive to the incentive scheme. We also take advantage of school climate surveys administered annually to teachers, principals 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.

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 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

²Lavy (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.

incentive programs show an even wider range of impact on student outcomes, with some finding negative results (Fryer, 2013; Goodman and Turner, 2013), others null impacts (Sojourner et al., 2014), and some positive, though typically modest, gains (Figlio and Kenny, 2007; Imberman and Lovenheim, 2015; Biasi, 2021). In all cases, extant evaluations of teacher incentives in the United States look only at test score gains, both immediate and in the longer-term.³

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, 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),

³The 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.

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. 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 over almost twenty 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 a minimum score in the first two domains to be eligible for additional awards. School administrators can also receive performance pay.⁴

Several comments on the incentive pay scheme under South Carolina TAP are warranted. First, there is

⁴The 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.

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 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.

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.

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

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 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 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

⁶Using 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.

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.

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 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.1 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 4 of Online Appendix Table A.1, 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 5).

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

⁸In 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.

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, 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.1).

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 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.

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

⁹Our 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.

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 intermediate outcomes (e.g., being held back in ninth grade) throughout the paper.

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. We show tabulations for the treated sample, by timing of TAP adoption, and 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).¹⁰ 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 an alternate comparison group: schools adopting the program outside our study window, 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.

¹⁰Composite standardized test score is the average of standardized test scores in ELA and math and is available for 33,459 students in our analysis.

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 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:

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

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 is the eighth grade cohort from the year

before TAP adoption, $\tau = -1$, and untreated units are included in this group as well. All other variables are as previously defined.

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).¹¹ 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 interaction weighted estimator of Sun and Abraham (2021). The Sun and Abraham estimator purges this source of bias 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). In its simplest form, the issue is related to the weighting

¹¹The 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.

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), as well as reproduce main findings using the imputation estimator proposed in Borusyak et al. (2021).

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.

Finally, Figure 1 depicts the cohort-specific point estimates by years elapsed relative to TAP implementation for key student outcomes. The height of the bars extending from each point represents the bounds of the 95% confidence interval. The lagged effects are generally small in magnitude and statistically indistinguishable from zero. 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.2 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.¹²

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 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).¹³ 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

¹²The 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. Because the tests and test scales changed dramatically, we limit our analysis in the last column of Table 3 to those students who were enrolled in fifth grade prior to 2008.

¹³Recall 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.

suggest any contamination due to attrition (Column 1 of Online Appendix Table A.4).

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,¹⁴ 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.

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 interaction weighted estimator of Sun and Abraham (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, the first column indicates that the implementation of TAP increased the probability of ever being enrolled in twelfth grade by a non-statistically significant 1.8 percentage points for the first post-adoption cohort, while the coefficient estimate for the fourth post-adoption cohort is 7.1 percentage points. The estimated effects for on-time graduation of the same cohorts are 0.026 (s.e.=0.021) and 0.078 (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,

¹⁴Note 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 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 (Borusyak et al., 2021).

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 Columns 3 and 4 are negative but fall short of statistical significance. Online Appendix Table A.3 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.3, continues to be smaller in magnitude and indistinguishable from zero. Online Appendix Table A.4 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.¹⁵

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 time since treatment and they also are more precisely estimated. For example, TAP adoption is associated with a 2.2 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

¹⁵Simple 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.4. Other arrests, reported in Column 5, are a heterogeneous group and include myriad offenses ranging from disorderly conduct to forgery.

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. The influence of TAP on economic self-sufficiency also continues to be more pronounced for later cohorts.

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.5 shows baseline estimates using the procedure (i.e., imputation estimator) outlined in Borusyak et al. (2021). The findings from this alternative approach are consistent with those presented throughout the text. Event studies using the same method further demonstrates that TAP implementation is not correlated with trends in long-run outcomes (Online Appendix Figure A.1).

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 correspond to 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.6). 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. The DiD estimates are larger in magnitude; however, we fail to reject the test of equality between these coefficients and those from baseline models (Table 4) across all the columns. Second, 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. 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.4).

Third, 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 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. Fourth, 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. Fifth, 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.

Sixth, 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. Seventh, we consider two alternate 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 adopters – 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.¹⁶

Eighth, we cluster the standard errors at the school-by-year level and such alternative clustering does not affect statistical significance. Finally, 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.6).

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 (Kling et al., 2007). As shown in the last column of Online Appendix Table A.4, 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.

In addition to these robustness checks, 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 .¹⁷ The results from this placebo analysis are reported in Online Appendix Table A.7.¹⁸ As expected, we find no

¹⁶We dropped the 2012-2013 academic year from the analysis when the comparison group is restricted to future adopters.

¹⁷We 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.

¹⁸As 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.

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. We also report the percentage of point estimates that are larger (smaller) than the baseline effects in Panels A and B (Panels C and D). 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.8 present 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.

We begin by presenting the impact of TAP on cognitive and non-cognitive high school outcomes in Table 5.¹⁹ 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). They also had higher composite standardized test scores in tenth grade. 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

¹⁹The available data for these intermediate outcomes vary. 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 gradually phased in as the state's high school exit examination and was fully administered for the first time in the 2005-2006 academic year. Students took HSAP in the spring of tenth grade. 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.

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.²⁰ The findings from a mediation analysis suggests that these intermediate outcomes can explain a sizeable fraction of the estimated impact of TAP on long-run outcomes.²¹

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 3). The results also provide suggestive evidence on TAP schools attracting lower quality and less experienced teachers, relative to school leavers, as evidenced by reduction in the fraction of teachers with advanced degrees and continuing contracts (Columns 2 and 4). This finding 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;

²⁰We 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). If anything, this biases our estimates in Columns 2 and 3 of Table 5 towards zero.

²¹To 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:

$$IO_{isc}^j = \alpha_1^j TAP_{sc} + X'_{isc} \alpha_2 + \delta_s + \lambda_c + \epsilon_{isc} \quad (3)$$

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

$$Y_{isc} = \beta_1^{Res} TAP_{sc} + X'_{isc} \beta_2 + \sum_j \theta^j IO_{isc}^j + \delta_s + \lambda_c + \epsilon_{isc} \quad (4)$$

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 $\frac{\alpha_1^j \theta^j}{\beta_1}$. For example, the decrease in the probability of being retained in ninth grade explains approximately 16 percent of the on-time graduation treatment effect, tenth grade composite test scores explain up to 10 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 15 percent of the on-time graduation effect.

Berlinski and Ramos, 2020). Note also that such compositional changes are likely to attenuate our results.

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, principals 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.

Finally, 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, ranging from raising returns to investment in schooling to changes in peer quality, or from increasing opportunity cost of crime to improved teacher quality (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.9 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.²²

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 18.²³ 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.²⁴ Panel A of Table 8 reports the

²²We 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.

²³Our 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).

²⁴We 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.

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 (\$9,932), which we proxy by expenditures per pupil averaged over 2006-2016, comes from Common Core of Data. 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.

7 Conclusion

Difference-and-differences and dynamic model estimates of the impact of TAP 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 through-

out 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 teacher incentives can be an effective school improvement strategy and that limiting evaluation outcomes to shorter-run outcomes may underestimate program effects. Given the program's effectiveness, it raises the question of which element of the program is most important for generating student success.

We note that we do not have the power to identify the impact of each TAP element separately. That said, a growing body of research casts doubt on the efficacy of teacher professional development programs 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, though the program effects may be driven by the incentives themselves or complementarity between the incentives and professional development. 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, *ming*).²⁵ These elements in and of themselves, or as a complement to teacher incentives and their particular design in this case, may be drivers of the TAP effect.

Given 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. A better understanding of the relative impacts of each of these design features is a useful area for future research.

²⁵Such teacher observations and feedback might be considered professional development. They stand in contrast to workshops and content-focused PD, which is the main form of PD evaluated in the surveys above.

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

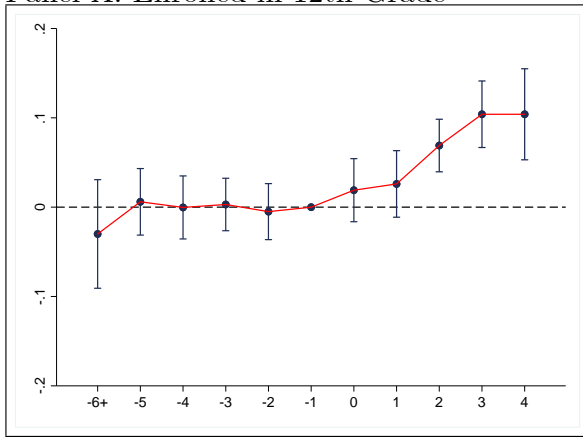
- Kling, J. R., J. Ludwig, and L. F. Katz (2005). Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment. *The Quarterly Journal of Economics* 120(1), 87–130.
- Lavy, V. (2020). Teachers' Pay for Performance in the Long-Run: The Dynamic Pattern 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.
- Muralidharan, K. and V. Sundararaman (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy* 119(1), 39–77.
- 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. Mykerezzi, 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.
- 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*.
- 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. (Forthcoming). Teacher evaluation and training. 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.

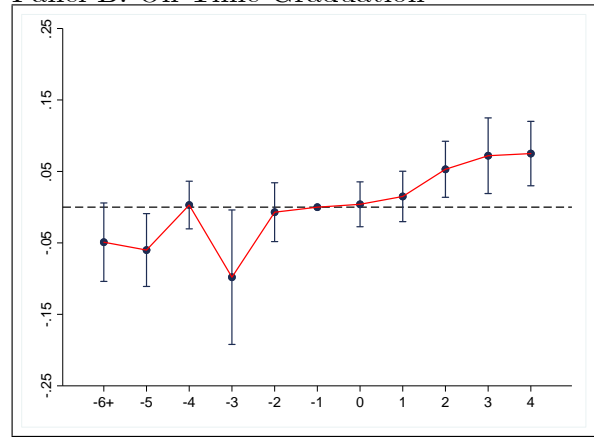
Tables and Figures

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

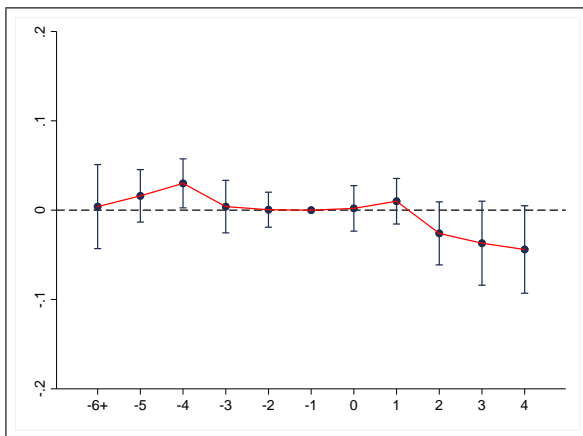
Panel A: Enrolled in 12th Grade



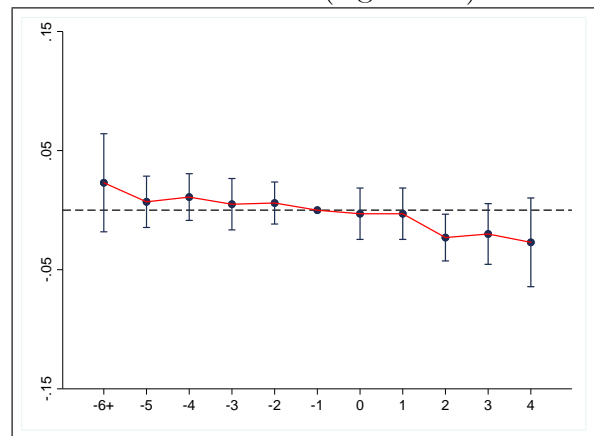
Panel B: On-Time Graduation



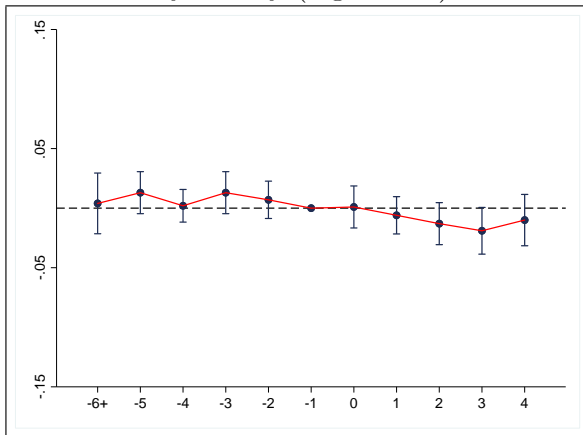
Panel C: Juvenile Crime



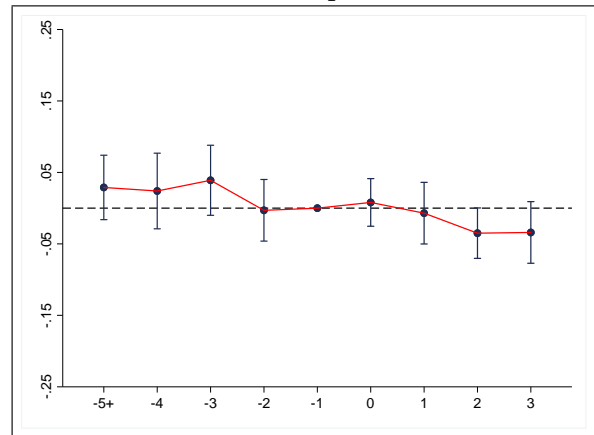
Panel D: Adult Crime (Age<=18)



Panel E: Any Felony (Age<=18)



Panel F: Welfare Receipt



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

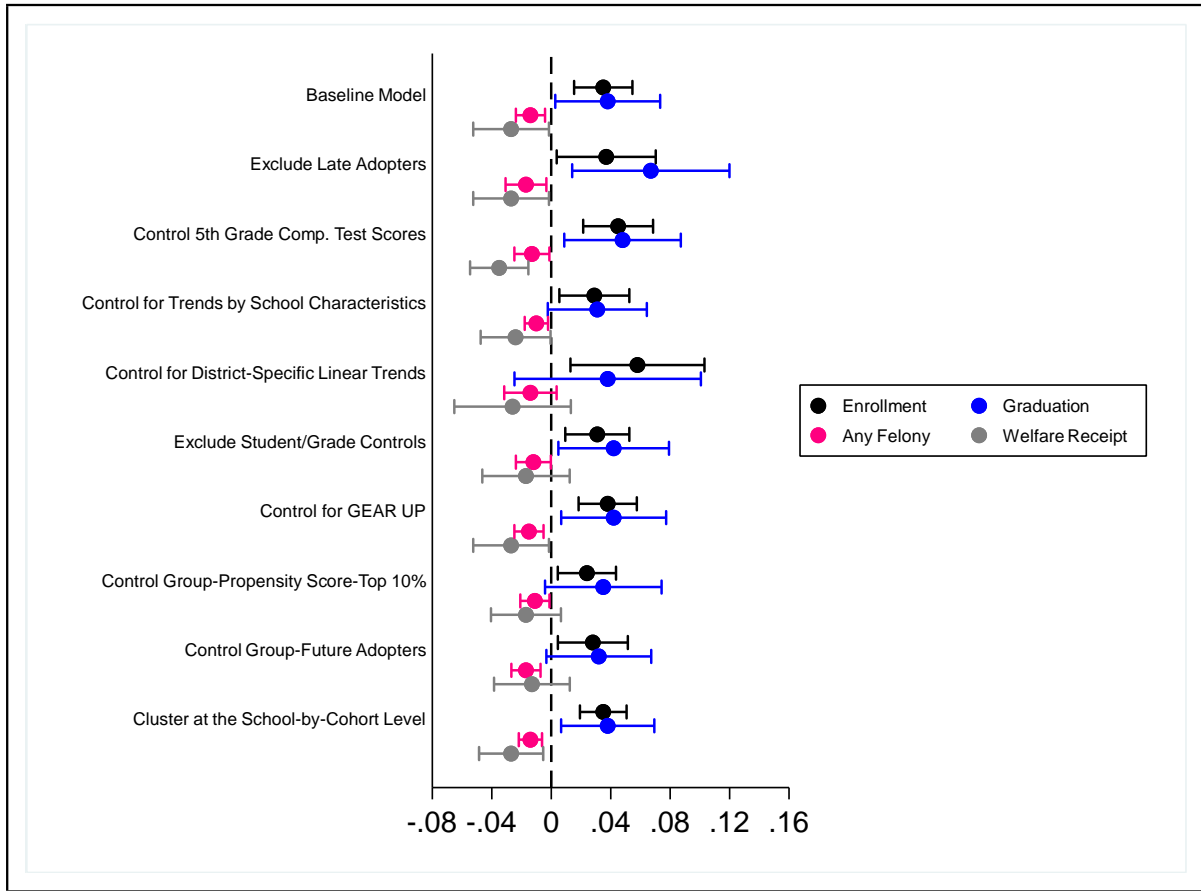


Figure 2: Effect of TAP on Long-Run Outcomes: Robustness Checks

Notes: Difference-in-differences estimates for models that exclude late TAP adopters, control for fifth grade composite test scores, control for trends by baseline school characteristics, control for district-specific linear trends, control for GEAR UP status of schools and use alternative control groups. Each row also shows 95% confidence interval based on standard errors clustered at the school level in Rows 1-8, while the last row displays confidence interval using standard errors clustered at the school-by-year level.

NOTES: Difference-in-differences estimates for models that exclude late TAP adopters, control for fifth grade composite test scores, control for trends by baseline school characteristics, control for district-specific linear trends, control for GEAR UP status of schools and use alternative control groups. Each row also shows 95% confidence interval based on standard errors clustered at the school level in Rows 1-8, 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

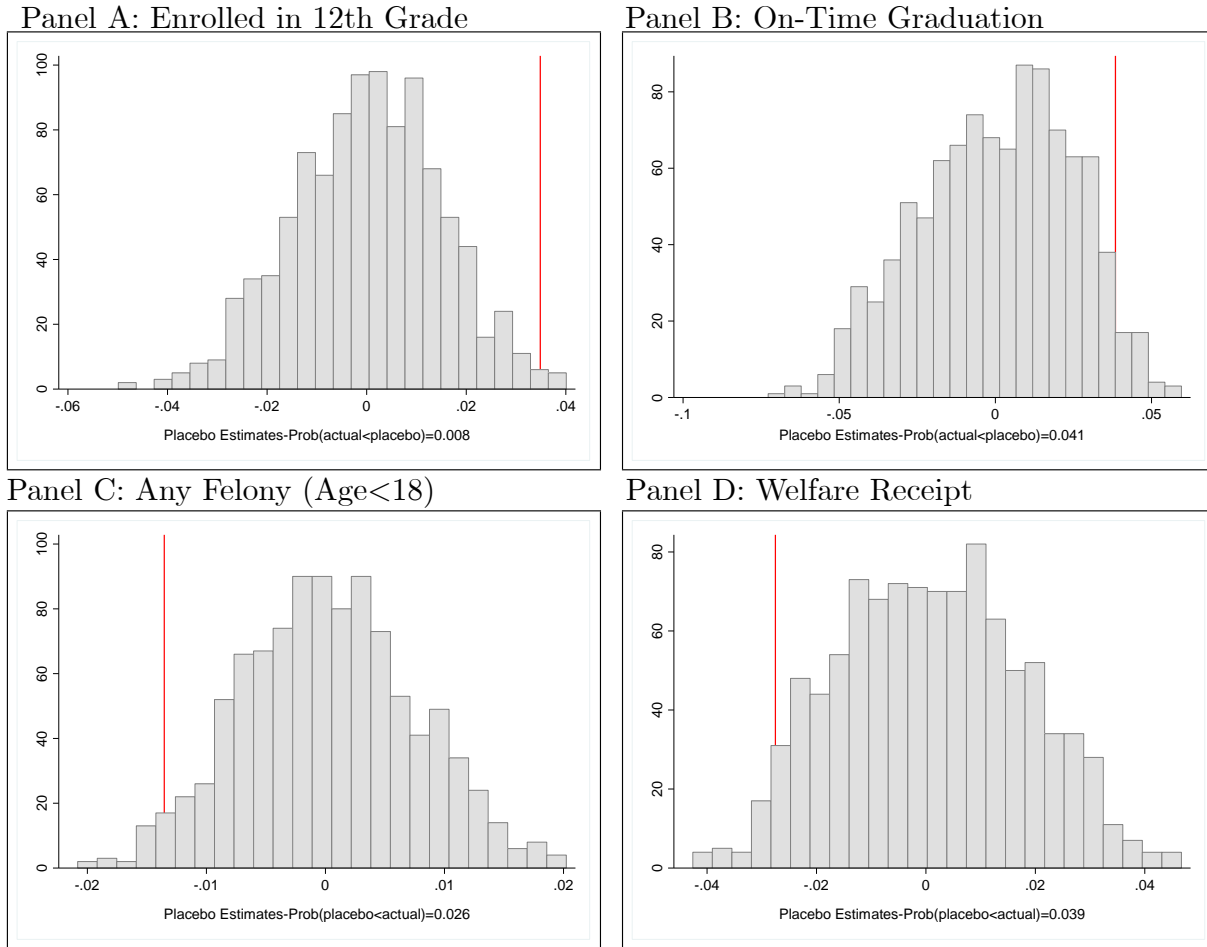


Figure 3: Placebo Estimates of the Effect of TAP on Long-Run Outcomes

Notes: 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. The fraction of placebo estimates that are greater (smaller) than the baseline estimates are reported on the x-axis of Panel A and B (Panel C and D). The fraction of placebo estimates that are greater (smaller) than the baseline estimates are reported on the x-axis of Panels A and B (Panels C and D).

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

	Trend	Dep Var: TAP Adoption Year
	(1)	(2)
Fraction of Female Students (8th Grade)	0.003 (0.004)	1,908.03 (5,354.96)
Fraction of Black Students (8th Grade)	0.006 (0.005)	-763.01 (1,458.15)
Fraction of Free/Reduced Lunch Students (8th Grade)	0.006 (0.005)	-2,388.68 (2,594.28)
Total School Enrollment	-1.620 (6.716)	13.119 (244.38)
Student Attendance Rate (%)	-0.011 (0.146)	24.949 (30.549)
Percent of Students Suspended/Expelled	0.070 (0.854)	1.652 (3.802)
Total Number of Teachers in the School	-0.454 (0.429)	-31.505 (62.424)
Percent of Teachers with an Advanced Degree	1.857 (1.571)	3.635 (15.761)
Percent of Continuing Contract Teachers	0.655 (0.621)	-8.421 (19.813)
Percent of Teachers Satisfied with Social and Physical Environment	-0.683 (0.945)	-12.791 (21.290)
Baseline (5th Grade) Composite Score	0.004 (0.013)	-430.39 (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 (1)	Black (2)	Free/Reduced Lunch (3)	Grade Size (8th Grade) (4)	Total School Enrollment (5)	5th Grade Test Scores (6)
TAP	-0.005 (0.014)	0.025 (0.018)	0.028 (0.018)	2.800 (8.262)	-32.181 (31.010)	-0.050 (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 \leq 18) (5)	Welfare Receipt (6)
Panel A: Difference-in-Differences Estimates						
TAP	0.035*** (0.010)	0.038** (0.018)	-0.012 (0.012)	-0.011 (0.007)	-0.014** (0.005)	-0.027** (0.013)
Panel B: Semi-Dynamic Model Estimates						
1st Postadoption Cohort	0.018 (0.013)	0.026 (0.021)	-0.008 (0.010)	-0.007 (0.008)	-0.010* (0.006)	-0.007 (0.015)
2nd Postadoption Cohort	0.025* (0.013)	0.038** (0.018)	-0.003 (0.011)	-0.006 (0.008)	-0.015** (0.006)	-0.025 (0.018)
3rd Postadoption Cohort	0.059*** (0.015)	0.066*** (0.023)	-0.029 (0.019)	-0.022** (0.010)	-0.018** (0.008)	-0.053*** (0.018)
4th Postadoption Cohort	0.071*** (0.022)	0.078** (0.039)	-0.027 (0.022)	-0.017 (0.011)	-0.014 (0.010)	-0.041* (0.021)
5th Postadoption Cohort	0.071*** (0.023)	0.081*** (0.031)	-0.033 (0.024)	-0.025 (0.016)	-0.022** (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 interaction weighted estimator from Sun and Abraham (2021). 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* (0.017)	0.066** (0.028)	-2.364 (2.185)
Comparison Mean	0.120	-0.085	18.22
Sample Size	40,800	27,123	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 in the School (1)	% of Teachers with Advanced Degrees (2)	% of Teachers Returning School from Previous Year (3)	% of Continuing Contract Teachers (Tenured) (4)
TAP	-0.389 (0.920)	-2.309 (1.903)	-3.481*** (1.172)	-3.823* (2.245)
Comparison Mean	31.64	54.48	82.21	66.67
Sample Size	302	302	302	302
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 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 1%.

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*** (0.155)	2.960 (1.956)	5.386** (2.222)	5.327*** (2.033)
Comparison Mean		78.53	73.17	74.10
Panel B: Students [N=293]				
TAP	-0.233 (0.148)	-2.905 (2.010)	-1.725 (1.970)	-2.014 (1.348)
Comparison Mean		71.91	74.28	81.43
Panel C: Teachers [N=294]				
TAP	0.099 (0.146)	3.361 (2.745)	2.206 (2.358)	-0.283 (2.811)
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
Teacher Performance Pay, Coaching, and Long-Run Student
Outcomes

Sarah Cohodes
Ozkan Eren
Orgul Ozturk

February 2023

Table A.1: Summary Statistics by School Type, Baseline (2002-2003) Academic Year

	TAP		All		Test of Equality		Test of Equality		Test of Equality	
	Schools	Comparison	Schools	Other	Cols. (1) vs. (2)	(p-value)	Cols. (1) vs. (3)	(p-value)	Cols. (1) vs. (4)	(p-value)
	(1)	(2)	(3)	(3)	(4)	(5)	(6)	(7)	(7)	(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 shows the alternative comparison group, schools that adopt TAP outside of the sample period.

Table A.2: 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 \leq 18) (3)	Welfare Receipt (4)
Panel A: All Districts				
TAP	-0.009 (0.014)	0.007 (0.018)	-0.006 (0.004)	0.005 (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 (0.014)	0.016 (0.018)	-0.007 (0.004)	-0.001 (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.3: **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** (0.004)	-0.009 (0.011)	-0.006* (0.003)	-0.011 (0.007)	-0.014 (0.012)
Panel B: Semi-Dynamic Model Estimates					
1st Postadoption Cohort	-0.005 (0.005)	-0.007 (0.010)	-0.004 (0.004)	-0.008 (0.007)	-0.009 (0.011)
2nd Postadoption Cohort	-0.010** (0.004)	0.001 (0.012)	-0.006* (0.003)	-0.003 (0.008)	-0.004 (0.013)
3rd Postadoption Cohort	-0.009* (0.005)	-0.026 (0.017)	-0.010* (0.006)	-0.021** (0.009)	-0.032* (0.019)
4th Postadoption Cohort	-0.009 (0.008)	-0.019 (0.021)	-0.005 (0.007)	-0.021** (0.009)	-0.025 (0.021)
5th Postadoption Cohort	-0.015** (0.007)	-0.031 (0.023)	-0.010* (0.006)	-0.024 (0.016)	-0.045* (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 interaction weighted estimator from Sun and Abraham (2021). 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.4: Effect of TAP on Additional Outcomes

	Juv/Adult Crime by Type				Composite Test		Long-Run Outcome Index (7)
	Attrition (1)	Violent (2)	Alc-Drug (3)	Property (4)	Other (5)	Score (5th Grade) (6)	
TAP	0.0004 (0.0041)	-0.009 (0.007)	-0.009** (0.003)	-0.009 (0.007)	-0.014 (0.010)	-0.027 (0.028)	0.047*** (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 several different 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 and 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 Table 4 and the text for further details. ** significant at 5%, *** significant at 1%.

Table A.5: **Effect of TAP on Long-Run Outcomes-Imputation Estimator (Borusyak et al. 2021)**

	Enrolled in 12th Grade	Graduated from HS in 4 Years	Any Juvenile Crime	Any Adult Crime	Any Felony (Age \leq 18)	Welfare Receipt
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Difference-in-Differences Estimates						
TAP	0.048*** (0.009)	0.058*** (0.023)	-0.019 (0.013)	-0.016** (0.006)	-0.015*** (0.005)	-0.027** (0.013)
Panel B: Semi-Dynamic Model Estimates						
1st Postadoption Cohort	0.021* (0.012)	0.030 (0.022)	-0.010 (0.009)	-0.008 (0.007)	-0.010** (0.005)	-0.004 (0.015)
2nd Postadoption Cohort	0.027* (0.014)	0.041** (0.019)	-0.004 (0.011)	-0.007 (0.007)	-0.015*** (0.005)	-0.023 (0.018)
3rd Postadoption Cohort	0.063*** (0.014)	0.073*** (0.024)	-0.031* (0.019)	-0.024*** (0.009)	-0.019** (0.007)	-0.049** (0.020)
4th Postadoption Cohort	0.077*** (0.019)	0.087** (0.039)	-0.030 (0.022)	-0.020** (0.010)	-0.013 (0.009)	-0.035* (0.020)
5th Postadoption Cohort	0.075*** (0.022)	0.086*** (0.032)	-0.035 (0.024)	-0.026* (0.015)	-0.021** (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 imputation estimator from Borusyak et al. (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.6: **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 (std) [sample size]	0.037** (0.017) [33,858]	0.067** (0.027) [29,997]	-0.017** (0.007) [33,858]	-0.027*** (0.013) [30,081]
Panel B: Control 5th Grade Comp. Test Scores				
TAP	0.045*** (0.012) [33,459]	0.048** (0.020) [33,370]	-0.013** (0.006) [33,459]	-0.035*** (0.010) [22,587]
Panel C: Control for Trends by School Characteristics				
TAP	0.029** (0.012) [43,062]	0.031* (0.017) [38,253]	-0.010** (0.004) [43,062]	-0.024* (0.012) [30,081]
Panel D: Control for District-Specific Linear Pretrends				
TAP	0.058** (0.023) [43,062]	0.038 (0.032) [38,253]	-0.014 (0.009) [43,062]	-0.026 (0.020) [30,081]
Panel E: Exclude Student and Grade Controls				
TAP	0.031*** (0.011) [43,062]	0.042** (0.019) [38,253]	-0.012** (0.006) [43,062]	-0.017 (0.015) [30,081]
Panel F: Control for GEAR UP				
TAP	0.038*** (0.010) [43,062]	0.042** (0.018) [38,253]	-0.015*** (0.005) [43,062]	-0.027*** (0.013) [30,081]
Panel G: Comparison Group-Propensity Score Top 10%				
TAP	0.024** (0.010) [56,335]	0.035* (0.020) [49,902]	-0.011** (0.005) [56,335]	-0.017 (0.012) [41,944]

Table A.6 – 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: Comparison Group-Future Adopters				
TAP	0.028** (0.012) [38,367]	0.032* (0.018) [34,017]	-0.017*** (0.005) [38,367]	-0.013 (0.013) [27,382]
Panel I: Clustering at the School-by-Cohort Level				
TAP	0.035*** (0.009) [43,062]	0.038** (0.016) [38,253]	-0.014** (0.004) [43,062]	-0.027** (0.011) [30,081]
Panel J: Wild Bootstrap				
TAP (p-value)	0.035** (0.016) [43,062]	0.038** (0.041) [38,253]	-0.014** (0.022) [43,062]	-0.027* (0.065) [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, while Panel B controls for 5th grade baseline composite test scores. Panel C adds interaction of baseline school level controls (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 with social and physical environment) with a linear trend. Panel D controls for district-specific linear pretrends. Panel E excludes student and grade level controls. Panel F conditions on schools' GEAR UP status. Panel G expands the matched control group and selects the top 10% of schools based on propensity scores. Panel H uses future TAP adopters, schools adopting TAP post-2012, as an alternate control group. Standard errors are clustered at the school level in Panels A-H, while they are clustered at the school-by-cohort level in Panel I. Panel J 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. See notes to Table 4 for further details. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table A.7: Placebo Effect of TAP on Long-Run Outcomes-Using Pre-TAP Adoption Years

	Enrolled in 12th Grade (1)	Any Felony (Age ≤ 18) (2)	Welfare Receipt (3)
TAP	-0.006 (0.014)	0.007 (0.005)	-0.001 (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.8: 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** (0.013)	0.034** (0.015)	0.041** (0.015)	0.021 (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** (0.017)	0.032 (0.023)	0.053** (0.022)	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 ≤ 18)				
TAP	-0.004 (0.004)	-0.023** (0.009)	-0.019** (0.008)	-0.009 (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* (0.013)	-0.031 (0.025)	-0.035** (0.016)	-0.011 (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%.

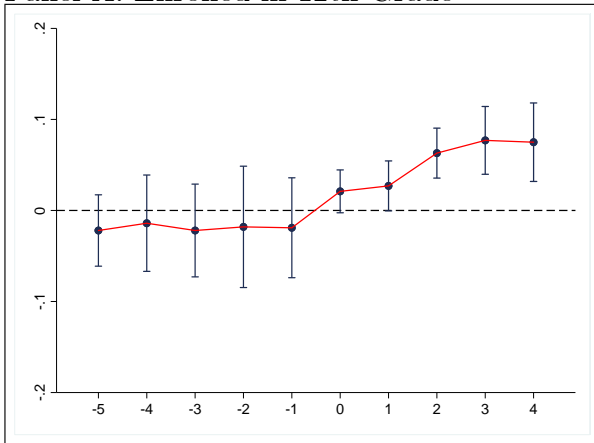
Table A.9: Effect of TAP on Long-Run Outcomes, Controlling for Ninth-Grade School Fixed Effects

	Enrolled in 12th Grade (1)	Graduated from HS in 4 Years (2)	Any Felony (Age \leq 18) (3)	Welfare Receipt (4)
TAP	0.033*** (0.010)	0.035** (0.017)	-0.012** (0.005)	-0.024* (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

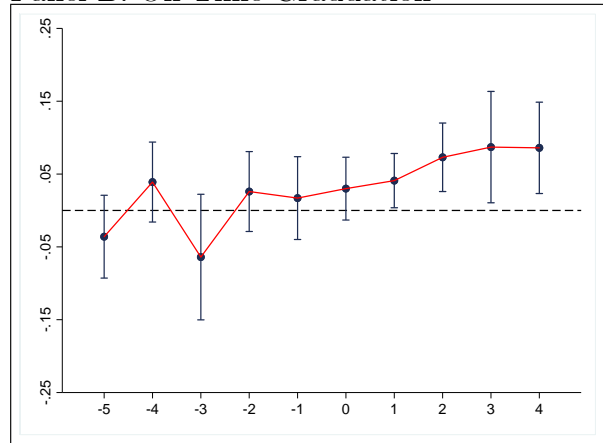
Notes: This table replicates the estimates from Table 4 but with ninth grade school fixed effects. Standard errors are clustered at the school level. All specifications include the same fixed effects and controls as the main specification. See Table 4 and the text for further details. ** significant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

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

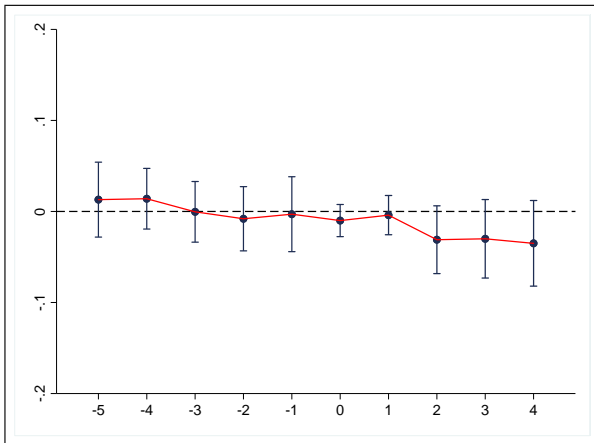
Panel A: Enrolled in 12th Grade



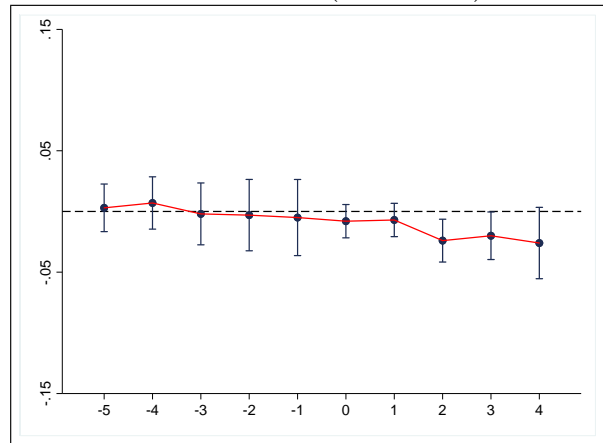
Panel B: On-Time Graduation



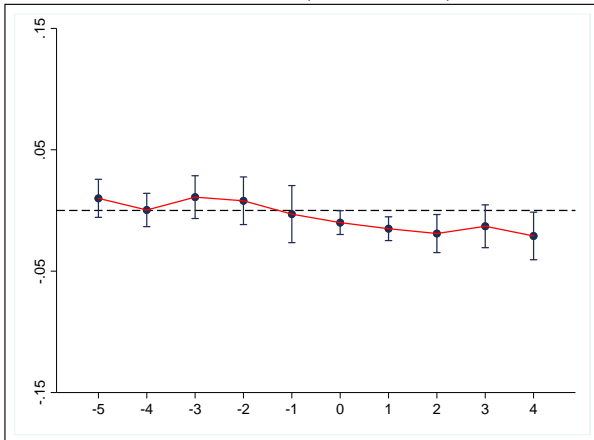
Panel C: Juvenile Crime



Panel D: Adult Crime (Age <= 18)



Panel E: Any Felony (Age <= 18)



Panel F: Welfare Receipt

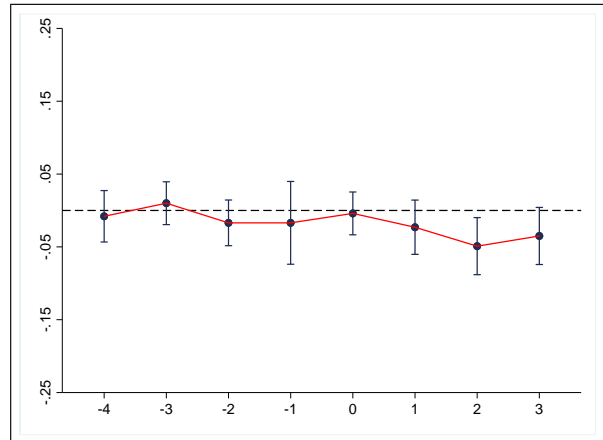


Figure A1: Event Study Estimates of the Effect of TAP on Long-Run Outcomes

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.