Journal of Public Economics 253 (2026) 105561

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/pubec

The long run effects of a teacher-focused school reform on student outcomes[☆]

Sarah R. Cohodes[a,b], Ozkan Eren[c,*], Orgul Ozturk[d]

[a] University of Michigan, Ann Arbor, MI, USA
[b] NBER, Cambridge, MA, USA
[c] University of California – Riverside, Riverside, CA, USA
[d] University of South Carolina, Columbia, SC, USA

HIGHLIGHTS

- Teacher-focused reform combining performance pay with teacher observation, and feedback improved student outcomes.
- Reform increased educational attainment, reduced felony crime activity, and lowered welfare reliance in early adulthood.
- Gains were driven by improved school climate and teacher efficiency, not sorting.
- Benefits exceeded costs.

ARTICLE INFO

Keywords:
School reform
Long-run outcomes
Difference-in-differences
Teacher incentives

ABSTRACT

This paper examines the effects of a teacher-focused school reform program — combining performance pay with teacher observation and feedback — implemented in high-need schools on students' longer-run educational, criminal justice, and economic self-sufficiency outcomes. Using linked administrative data from a Southern state, we leverage the quasi-randomness of the timing of program adoption across schools to show that the school reform improved educational attainment and reduced both felony criminal activity before age 19 and dependence on government assistance in early adulthood (ages 18–22). We find little scope for student sorting or changes in the composition of teacher workforce to explain the findings, and instead find changes in school climate consistent with improved school environments and increased teacher efficiency. Program benefits far exceeded its costs. A comparison with a similar educator-focused reform suggests that the individual incentive component of the program is necessary but not sufficient to improve student outcomes.

1. Introduction

Improving low-performing schools is a perennial problem in education systems. Policymakers have implemented many strategies to turn around struggling schools, with varying degrees of success. One promising possibility is the use of teacher incentives. While performance pay may increase costs, it is unlikely to necessitate a large-scale hiring of staff or a rehauling of school curricula 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 et al., 2016; Schueler et al., 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 studies of such incentive programs have

[☆] We are grateful to Marianne Bitler, Katharine Parham Malhotra, Richard Mansfield, Jonah Rockoff, Eric Taylor, Yotam Shem-Tov and seminar participants at Vanderbilt University, the University of South Carolina, Clemson University and the University of Central Florida, the 2023 conference of the Society of Labor Economists and the 14th Workshop on the Economics of Risky Behavior 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.

[*] Corresponding author.
Email addresses: scohodes@umich.edu (S.R. Cohodes), ozkane@ucr.edu (O. Eren), odozturk@moore.sc.edu (O. Ozturk).

https://doi.org/10.1016/j.jpubeco.2025.105561
Received 3 March 2025; Received in revised form 28 November 2025; Accepted 23 December 2025
Available online 2 January 2026

been limited to test score outcomes—measures that may not fully encompass the impacts of teachers and teacher effort (Imberman, 2015; Jackson, 2018). By contrast, growing evidence shows that interventions such as teacher observations and feedback can improve student performance (Taylor and Tyler, 2012; Briole and Maurin, 2022; Taylor, 2023). Therefore, a teacher-focused school reform program that combines *both* incentives and support through teacher observation and feedback may address some of the limitations of narrower pay-focused initiatives. In analyzing the efficacy of an educational intervention and the role of teachers more broadly, it is further important to gauge whether short-term effects (if any) persist or fade out and whether these programs spur meaningful change in long-run outcomes. An informed debate is crucial to designing optimal public policies.

In this paper, we study the medium- and long-run effects of a teacher-focused school reform 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 comprehensive school initiative, which embeds incentive pay 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 provide 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 such as 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), as well as teacher performance pay programs introduced in several North Carolina school districts, evaluated in Hill and Jones (2020), resulted in no and sometimes negative test score impacts. Again, the incentive scheme in these interventions contained little guidance on how to improve student performance.

In addition to its comprehensive nature, incentive pay for teachers under the TAP system differs from other incentive systems in three important ways. First, teachers' bonus allocations hinge on both their own students' achievement gains as well as the school's overall achievement growth. In this regard, TAP is a hybrid program involving both individual and group incentives, and is thereby less likely to suffer from the challenges facing either pay scheme in isolation—foregone benefits under individual incentives, due to lack of cooperation, or free-riding under group incentives. (Holmstrom, 1982; Muralidharan and Sundararaman, 2011). Second, bonuses are substantial and individualized, and thus may be more likely to cause changes in behavior than egalitarian distribution methods that weaken individual incentives (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. For example, measuring performance solely by student performance on standardized assessments may encourage teaching to the test or crowd-out promotion of higher-order skills (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 dozens of 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. Since the majority of TAP schools are high-need, we rely on propensity score matching to identify a set of comparison schools that are most similar to TAP schools prior to the implementation. Our identifying assumption is that any unobserved factors influencing outcomes, such as a different school reform policy, would have evolved similarly in TAP and comparison schools. For any observed differences in outcomes to be driven by such 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 showing little scope for such changes in comparison schools to undermine our conclusions, including tests for the existence of pretrends and endogenous mobility, conditioning on district-specific trends, experimenting with alternate comparison samples and inference approaches, controlling for potential concurrent policy changes, and excluding schools with varying grade configurations. We also conduct placebo tests which support this identifying assumption. We employ the procedure in Borusyak et al. (2021) to address concerns about different forms of biases in settings with staggered treatment rollout and further complement these results using the interaction weighted estimator of Sun and Abraham (2021).

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' felony arrest rates before age 19. Specifically, students in TAP schools were 1.4 percentage points less likely to be arrested for a felony offense post-program adoption (a 30 percent decrease relative to the comparison mean). We also find negative but smaller and less precisely estimated effects of TAP on non-felony offenses. Finally, the TAP program decreased the odds of reliance on Supplemental Nutrition Assistance Program (SNAP) and Temporary Assistance for Needy Families (TANF) 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. Gains in high school outcomes can account for about half of the observed change in longer-term outcomes. 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, equivalent to one teacher per year from an average of 32 teachers. Further examination provides evidence of TAP schools attracting less educated, less qualified, and less experienced teachers relative to teachers who left. If anything, these changes in teacher workforce composition, following the teacher effectiveness literature, should worsen student outcomes. This implies that teacher sorting does not account for the program's benefits. Instead, evidence from school climate surveys administered annually to teachers, students, and parents implies that TAP changed the school experience. We find that the fraction of parents and teachers who are satisfied with learning and social and physical environments increased in the post-adoption period, although the effects for teachers are less precisely estimated. Students are not more satisfied, perhaps due to the additional effort asked of them. Taken together, our findings are consonant with explanations related to improvements in school climate as well as increases in the productivity of incumbent teachers.

In an attempt to isolate the effects of components of this bundled intervention, we contrast South Carolina TAP with a version implemented in Chicago (see Glazerman and Seifullah, 2012). The test-based individual teacher incentive pay, originally integrated into the program, was omitted in Chicago TAP because of lack of reliable data for linking students and teachers. The programs were otherwise identical, such as the observation and feedback mechanisms for teachers, yet impacts of the programs diverged. Chicago TAP did not boost student test scores, whereas, as we show, South Carolina TAP did (as well as improving the longer-term outcomes we focus on here). This comparison suggests that individual incentives are a key component of the TAP intervention, though such incentives may need to be embedded in a comprehensive program to be effective.

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 SNAP/TANF receipt 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 four 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, bundled with teacher support, may shape students' longer-term outcomes, rather than just test scores.[1] Evidence on teacher incentive programs shows wide range of impacts on student outcomes, with some finding negative results (Fryer, 2013; Goodman and Turner, 2013; Hill and Jones, 2020), some positive, though typically modest, gains (Figlio and Kenny, 2007; Sojourner et al., 2014; Springer et al., 2014; Dee and Wyckoff, 2015; Imberman and Lovenheim, 2015; Eren, 2019; Biasi, 2021; Hanushek et al., 2023; Morgan et al., 2023), and others null impacts (Glazerman and Seifullah, 2012; Hill and Jones, 2020).[2] Our findings suggest that a comprehensive pay scheme, embedded with observations of teaching practices and a feedback mechanism, can deliver desired student outcomes in a cost-effective way.

Second, we add to the evidence on school turnaround strategies more broadly. TAP was targeted at low-performing schools, which are a frequent subject of education reform efforts designed to increase student performance. Such efforts include comprehensive school reform (Borman et al., 2003, 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 include 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 takeovers 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 at improving school performance, but one that works with existing school staff and practices, focusing on teachers to improve student performance.

Third, our findings add to the literature on the causal effects of education on crime (Lochner and Moretti, 2004; Deming, 2011; Hjalmarsson et al., 2015; Cook and Kang, 2016; Davis and Heller, 2020; Jackson et al., 2020; Rose et al., 2022; Alsan et al., 2024). This literature generally shows that human capital accumulation or access to better schools deters crime. Like Deming (2011) and Jackson et al. (2020), the TAP gains found here suggest that higher-quality schooling deters criminal involvement.

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 SNAP/TANF receipt). 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, 2020, for overviews and discussions of this phenomenon). For example, researchers have shown that short-run effects do not fully capture long-run effects when examining Head Start and other preschool programs (Ludwig and Miller, 2007; Gray-Lobe et al., 2021; Anders et al., 2023), class size (Chetty et al., 2011; Dynarski et al., 2013), school choice (Deming et al., 2014; Beuermann and Jackson, 2022), accelerated learning (Cohodes, 2020), and Medicaid access for children (Cohodes et al., 2016). Our findings that modest test score gains precede meaningful increases in educational attainment, decreases in criminal activity, and decreases in reliance on SNAP or TANF 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

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. The National Institute for Excellence in Teaching (NIET), an independent public charity, manages the nationwide implementation of TAP.

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

The second element of TAP, ongoing applied professional growth, allows teachers to learn new instructional strategies, collaborate with master and mentor teachers, and receive individual coaching. Teachers meet in grade-alike or subject-alike groups under the guidance of master and mentor teachers for about 50 to 90 minutes each week. Instruction-focused accountability, the third program component, requires teachers in TAP schools to be held accountable for high-quality instruction.

[1] Lavy (2020) examines the effects of a performance-based compensation program for teachers, which was conducted in 49 Israeli high schools, on long-term human capital and labor market outcomes. 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.

[2] The evidence on the impact of incentive pay on student achievement from other countries is more encouraging. See, for example Lavy (2002) for Israel; Atkinson et al. (2009) for England; Glewwe et al. (2010) for Kenya; and, Muralidharan and Sundararaman (2011) for India.

Teachers are evaluated four to six times during the school year by school administrators and master and mentor teachers in different areas of effective instructional practice for an overall classroom observation score. Post-evaluation sessions are also held by observers to help teachers strengthen their instructional practices. Finally, teachers in TAP schools are eligible for additional compensation based on their performance in the classroom (observations of 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. Schools were selected by NIET from among those that demonstrated the capacity to implement the program (Institute of Education Sciences, 2015). As discussed in Section 3.2, TAP's staggered adoption in South Carolina forms the basis of our identification strategy.[3]

South Carolina TAP implemented the accountability and performance compensation aspects of TAP via a formula that weighted classroom observation scores, individual value-added scores (for teachers in relevant subjects), and school-level value-added scores. 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. 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 may be eligible for additional awards based on high individual rankings within their school on classroom observation and individual value-added scores. For teachers in grades and subjects in which state assessments are not administered, bonus allocation is based on school achievement growth and teacher observations (evenly weighted). Teacher leaders also received special pay under TAP, which was a mix of pay-for-services and incentive pay. Over the analysis period, for their additional duties, NIET recommended annual compensation of $5000 to $8000 for mentor teachers and $8000 to $12,000 for master teachers. School administrators also received performance pay based mostly on school-level value-added scores and these bonuses ranged from $0 to $14,000 for the 2009–2010 academic year (South Carolina Department of Education, 2012).[4]

Several comments on the incentive pay scheme under South Carolina TAP are warranted. First, there is no consensus on how to design optimal teacher incentives (Jackson and Bruegmann, 2009; Muralidharan and Sundararaman, 2011; Fryer, 2013; Goodman and Turner, 2013; Imberman and Lovenheim, 2015; Brehm et al., 2017). While it is conceivable that individual incentives dominate group-based incentives because of the free-riding problem inherent in group incentives, complementarities and gains to cooperation may ultimately make group-based incentives a more effective tool. South Carolina TAP is a hybrid program involving both individual and group incentives and thus it is less likely to suffer from the design-specific concerns of simpler incentive pay schemes. Second, bonuses were substantial and sufficiently differentiated to cause changes in the behavior of educators. For example, the average incentive pay for teachers across the state was approximately $2,000, ranging from $0 to $10,000, for the 2009–2010 academic year (South Carolina Department of Education, 2012). The average teacher salary in that 2009–2010 in South Carolina was $47,508. As such, the average incentive pay was equivalent to four percent of the average teacher salary, with maximum incentive pay equal to greater than 20 percent of the average teacher salary.

Third, incentive pay was not solely determined by teaching outcomes. Observations of teaching practices, which were coupled with professional feedback indicating how to improve performance, 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] The incentive scheme was convex, with higher within-school ranking resulting in greater performance pay: it was possible for the highest ranked teachers to earn five times more than the average bonus. In this respect, the structure of the bonus pay includes both absolute targets and rank-order tournaments and does not necessarily imply egalitarian distributions where an overwhelming majority of teachers receive the same award. Appendix B provides details of TAP compensation using a hypothetical example.

The complexity of the incentive scheme itself—multiple measures, which differ in their weights depending on teacher subject, as well as use of value-added scores—is further compounded by complicated formula by which it is possible to achieve different levels of bonuses based on within-school ranking (Appendix B), This may also influence teachers' responses to TAP. One possibility is that complex incentive schemes attenuate behavioral responses, as complexity can impede decision-making in multiple contexts (taxation: Abeler and Jäger (2015); take-up of social programs: Kleven and Kopczuk (2011); financial aid: Dynarski and Scott-Clayton (2006); and retirement contributions: Choi et al. (2009)). However, recent empirical work has shown that incentive complexity can sometimes increase effort since workers overproduce when they do not understand the exact incentive scheme (Abeler et al., 2023). There is also evidence on how the salience of incentives (even complex schemes) matters (Englmaier et al., 2017). The most salient aspect of the TAP incentive scheme to teachers is the observation and feedback component, as it is a change to their daily experiences in school. The observation rubric is known by teachers, and observers are trained colleagues who also conduct a pre- and post-observation conference. The rubric primarily focuses on domains directly related to the student learning experience: instruction, planning, and environment; with one additional performance area: professionalism. This is very similar to the Danielson rubric used in Cincinnati and shown to be effective at increasing teacher value-added in an observation and feedback framework (Taylor and Tyler, 2012). Thus while the incentive scheme itself is complex, teachers' experiences include direct, salient feedback on areas to improve performance and thus increase their incentive payout.

[3] The investigation of the long-run effects of TAP exposure on socioeconomic outcomes entails focusing on student population who will be in their early adulthood by the end of our sample period (i.e., 2017). As a result, our analysis sample only includes TAP schools whose grade configuration contains eighth grade (K-8, 3–8 and all middle schools). Such sample restriction does not pose any threat to causal identification as grade configuration is not endogenously determined by TAP. Nevertheless, we provide an extended analysis of the program by including elementary TAP schools in Online Appendix C.

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

[5] On average, each TAP school allocates $2000 to $3000 per teacher to establish the award pool (Institute of Education Sciences, 2015).

3. Data and methods

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-price lunch status and age, test scores from selected grades and information on high school graduation. In addition, for a subset of academic years, we have records of attendance for each student. Unique identification numbers allow us to track all the students through their tenure in the public school system from the fall of 2000 to the spring of 2017. The SCDOE data do not include information on individual teachers. It is thus not possible to link students to teachers.

The juvenile crime data come from the South Carolina Department of Juvenile Justice (SCDJJ) and include the universe of detailed arrest records from 2000 to 2017. For each juvenile offender file, we have basic demographic information on the arrestees, offense date and the type of crime they are arrested for. We complement these data by drawing information on administrative records from the South Carolina State Law Enforcement Division (SC 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 some government assistance programs, which is available through 2019. We are able to link individuals' records across these four data sets. Any students who do not have a match are coded as zeros for their respective long-run socioeconomic outcome variables.[6] 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/home-schooling 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. Using the American Community Survey data, we find that less than 7 percent of the population born in South Carolina in 1990s left the state at age 18 or earlier.

3.1.2. 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. We focus on four key indicators to summarize students' well-being as they enter adulthood: twelfth grade enrollment, on-time high school graduation, felony crime arrests, and receipt of some forms of government aid.

Measures of educational attainment include twelfth grade enrollment status and on-time high school graduation.[7] These are the only educational attainment variables available in the South Carolina data. On-time high school graduation is a relatively narrow indicator of educational attainment. This is because it misses students who graduate later or through alternative pathways. Since virtually all students who graduate must first enroll in twelfth grade, we also report the TAP effects using twelfth grade enrollment as a proxy for the broader high school graduation margin.

Records from the SCDJJ and SC SLED allow us to examine criminal activity. The criminal justice outcomes available to us are arrest records, separated by juvenile (age 16 and under) and adult ages. To follow all cohorts for the same time period, we censor adult arrests to those prior to age 19.[8] The arrest records also include crime severity (felony and non-felony) and crime type. We construct different measures of crime by age at arrest, severity, and crime type. Our preferred crime outcomes combine juvenile and adult criminal engagement prior to age 19 separately for felony and non-felony offenses. The combination of juvenile and adult criminal arrests reflects the well-documented age–crime pattern in the U.S.: criminal involvement rises sharply during adolescence, peaks in the late teenage years, and declines rapidly thereafter (Levitt and Lochner, 2001).

In line with current practice in the literature (Deming, 2011; Aizer and Doyle, 2015), we focus primarily on felonies for three reasons. First, felonies impose substantially greater social costs than misdemeanors, making them more policy-relevant margin of criminal activity. Second, the long-run consequences of a felony conviction are far more severe than those of a misdemeanor (Agan et al., 2024). Finally, potential reporting bias is far less of a concern for felonies. For example, TAP schools may have incentives to underreport less serious infractions, and such underreporting may even persist into non-TAP years if schools tend to underreport for students who are performing better academically.

Records from the SCDSS allow us to construct a 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 that 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. Receipt of SNAP and TANF may reflect a lack of need for such government aid, if young adults are more likely to have sufficient means without such transfers. Alternatively, lack of receipt may reflect difficulty accessing benefits or being removed from aid due to changes in eligibility. We cannot distinguish between these possibilities with our data, though we note that increases in education and decreases in criminal activity are more likely to be consonant with the former explanation than the latter. It is worth mentioning that these conditional cash and in-kind transfers constitute an important source of income for recipients in South Carolina. Using the 2010–2019 SNAP Quality Control files provided by Mathematica Policy Research, Inc., we find the average monthly SNAP benefit ($210 in 2015 dollars) to be roughly equal to 20 percent of the total gross income recipients reported.

Finally, to perform a comprehensive evaluation of TAP and explore various mechanisms, we also consider several shorter-run outcomes (e.g., being held back in ninth grade, mandatory high school exit exams taken in the spring of tenth grade) throughout the paper. The tests and test scales administered in elementary and middle schools changed dramatically beginning with the 2008–2009 academic year which prevents us from analyzing the efficacy of the program on eighth (and earlier) grade test scores. The change was made in an effort to provide a more comprehensive assessment of student learning and ensure that the state's standardized testing program is in line with current educational standards.[9]

[6] We cannot determine the match rate from SCDOE to other registers since it is not feasible to establish a linkage for individuals without a record in crime or welfare data.

[7] Our on-time graduation analysis excludes eighth graders from the 2002–2003 academic year because SCDOE provided information on graduation beginning with the 2007–2008 academic year.

[8] Online Appendix Table A.1 shows outcome availability by cohort.

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

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 High School 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 (age 17–18)	0.086	0.095	0.072	0.068	0.066
Any Felony (age ≤18)	0.056	0.062	0.046	0.045	0.040
SNAP/TANF Receipt (age 18–22)	0.514	0.527	0.498	0.607	0.548
Sample Size	29,645	17,761	11,884	13,417	9575

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 using the top 5 % of schools in terms of similarity as determined by propensity score matching. 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.

3.1.3. 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. We focus on 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 and these are the students old enough to achieve longer-term outcomes. This results in 21 TAP adopting schools in the sample. Online Appendix Table A.1 shows the implementation years of TAP and associated outcome years.

Given that the majority of TAP adopters are high-need schools serving large fractions of disadvantaged students, one would expect TAP schools to be different from 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 that are most similar to TAP schools in terms of observable characteristics prior to the adoption of TAP. 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 add other school characteristics that we believe should be part of the propensity score model. Online Appendix Table A.2 presents these school characteristics from the baseline academic year.

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

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

3.1.4. Descriptive statistics

Columns 1–4 of Table 1 present descriptive statistics for a total of more than 43,000 students from 31 unique schools. Online Appendix Figure A1 shows the distribution of grade configurations for these schools based on the highest grade offered. We show tabulations for the treated sample, by timing of TAP adoption, and for a matched comparison sample. As displayed in Panel A, Black and White students comprise 53 and 43 percent of all students in TAP schools, respectively, and 67 percent of the treated sample received free/reduced-price lunch (Column 1). Students in matched comparison schools are more likely to be Black, come from disadvantaged families, and have lower baseline composite test scores (Column 4).[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 for 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.

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

Finally, the last column shows the same descriptive statistics from our main alternate comparison group, the so-called "future adopters." This alternate sample is very similar to that from Column 2 in observable student characteristics, but the sample size is almost two-thirds of our preferred comparison group since we remove the 2012–13 cohort from the analysis sample. As noted above, the similarity of the estimated impacts of TAP from alternate comparison groups may provide assurance as to the credibility of the identification strategy.

3.2. Empirical methodology

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

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

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, as well as 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 students exposed at different times for different lengths of time, β_{DiD} represents any TAP exposure and is a weighted average of each of the cohort effects 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 that is easy to compare across multiple specifications. However, to investigate dynamic response to treatment, we also estimate flexible event study specifications of the following form:

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

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 cohorts from 6 or more years before TAP adoption, $\tau = -6^+$, and untreated units are included in this group as well. All other variables are as previously defined. In Section 4.3, we demonstrate that the estimates from the event study model remain robust to alternate choices of the event time window, revealing that binning the pre-endpoints does not pose any threat to identification.

Treatment effects that occur in response to TAP and vary over time are indicated by γ_0 to γ_4 and trace out impacts on student outcomes by cohort of exposure to TAP. For example, in a school with grade 6–8 configuration, students in the initial exposure cohort will be in a TAP school for a single year (eighth grade). Students in the next cohort are typically exposed to TAP for two years (seventh and eighth grade), and students in the next and subsequent cohorts are exposed for three years (sixth, seventh, and eighth grades).[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 reassuring 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 experiencing 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 Eq. (2) using the imputation estimator of Borusyak et al. (2021).

The imputation estimator purges this source of bias by generating predicted values of the outcome for students in TAP schools in the post-adoption period using the two-way fixed effects model described above for only the non-treated observations (students in non-TAP schools and yet-to-adopt TAP schools). An estimate of the treatment effect can then be obtained for each treated observation by calculating the difference between their observed and predicted outcome (in both pre- and post-adoption periods for event study models) and taking the average of these differences. As a further robustness check, we estimate the event study models using the interaction weighted estimator of Sun and Abraham (2021).[12]

Another important threat to identification in settings with variation in treatment timing stems from the negative weighting problem (Borusyak et al., 2021; Goodman-Bacon, 2021). In its simplest form, the issue is related to the weighting scheme implicit in OLS where the weights of the dependent variable are proportional 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).

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

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

Table 2
Trends in school characteristics prior to TAP adoption and predicting TAP adoption year.

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

Notes: Each cell in Column (1) presents a separate regression where the key coefficient of interest is on a trend in the number of years since TAP adoption. The regression specifications, which control for cohort and school fixed effects, include indicators for each post-adoption year and therefore, the point estimates in Column (1) can be interpreted as a test for whether there is a significant pre-trend for each outcome. Column (2) tests whether the year of TAP adoption is associated with school characteristics from the baseline (2002-2003) academic year. Standard errors are clustered at the school level in Column (1), while heteroskedasticity-robust standard errors are reported in Column (2). The F-test p-value comes from a test that the coefficients shown are jointly equal to zero. * significant at 10 %, ** significant at 5 %, *** significant at 1 %.

Prior to continuing, it is worth mentioning that this negative weighting problem is inherently different from 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 (ii) there are no coincident shocks or policy adoptions that could account for the TAP effect. We provide three sets of evidence supporting the plausibility of the first assumption. First, we test whether TAP adoption was preceded by a systematic change in school characteristics. To diagnose the importance of any pre-existing trend, we estimate a modified event study by replacing the pre-TAP indicators with a linear trend. The parameter of interest in this specification yields the slope of school characteristics over time prior to TAP adoption. The first column of Table 2 presents these coefficient estimates. Of the 11 outcomes we analyze, none is statistically significant at even the 10 % level. Second, we examine the associations between the year of TAP adoption and baseline school characteristics. As shown in Column 2, the covariates do not significantly predict the timing of TAP adoption. The p-value for joint significance is 0.56. The inference on the predictive power of the covariates does not change when we replace the year of adoption with indicator variables for being an early or late adopting school, as reported in Online Appendix Table A.3.

Finally, Fig. 1 depicts the cohort-specific point estimates by years elapsed relative to TAP implementation for key student outcomes. The length of the bars extending from each point represents the bounds of the 95 % confidence interval. The lagged effects are generally small in magnitude and individually statistically indistinguishable from zero. The pre-adoption coefficients are also jointly equal to zero across all panels.[13] Taken together, we see no trends in educational attainment, crime and self-sufficiency outcomes from cohorts in TAP schools prior to TAP adoption. The second assumption is not directly testable. However, we show in Online Appendix Table A.4 that when we characterize TAP implementation at the district level, with TAP adoption beginning when any school in the district introduces TAP program, we find no impacts on student outcomes. This implies that TAP is not part of some larger package of district-level programming adopted at the same time.

Next, we test for post-adoption student sorting. Table 3 shows the impact of TAP exposure on eighth grade student characteristics. Neither the proportion of girls, Black students, nor students who received

[13] The corresponding p-values in Fig. 1 for joint significance are 0.84 in Panel A, 0.33 in Panel B, 0.63 in Panel C and 0.33 in Panel D.

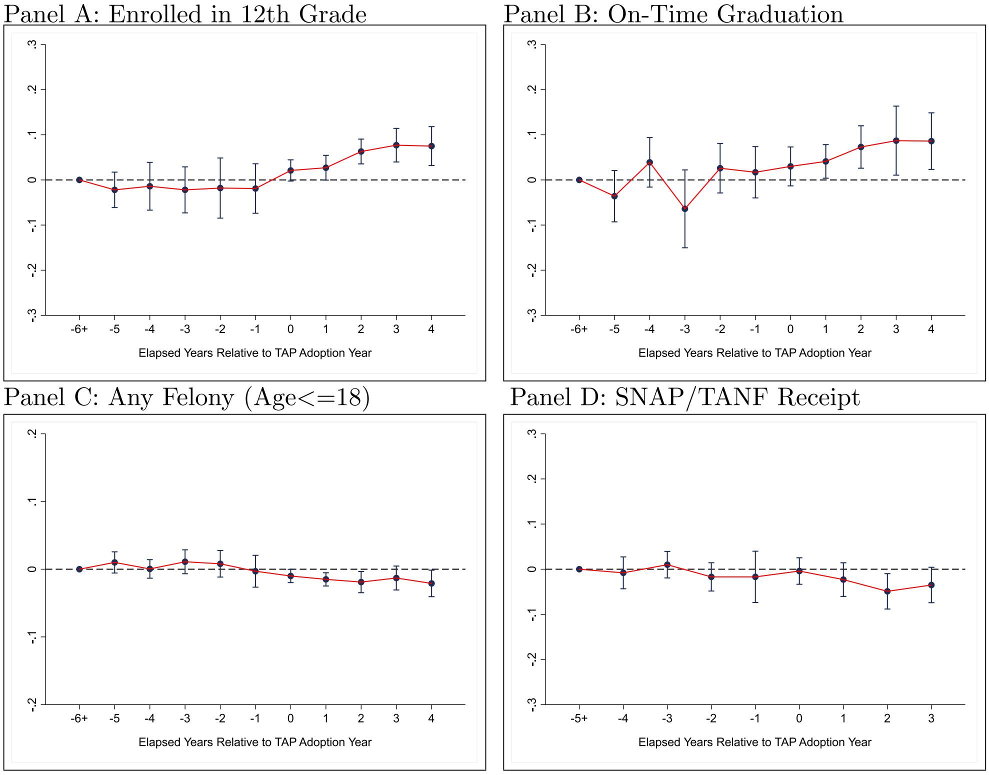


Fig. 1. Event study estimates of the effect of TAP on long-run outcomes. Notes: This figure shows event study estimates for various outcomes obtained using the imputation estimator from Borusyak et al. (2022). Each panel shows coefficient estimates and 95 % confidence intervals based on standard errors clustered at the school level. Six or more years before TAP adoption ($\tau \leq -6$) is the omitted category.

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	0.025	0.028	2.800	−32.181	−0.050
	(0.014)	(0.018)	(0.018)	(8.262)	(31.010)	(0.035)
Sample Size	302	302	302	302	302	230

Notes: This table reports difference-in-difference estimates of TAP exposure on the school characteristics listed in the column headings. 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. * significant at 10 %, ** significant at 5 %, *** significant at 1 %.

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 in prior test scores.[14] As noted in Section 3.1, because the tests and test scales changed dramatically beginning with the 2008–2009 academic year, we limit our analysis in the last column of Table 3 to those students who were enrolled in fifth grade prior to 2008.

Finally, we examine whether TAP adoption is correlated with sample attrition. Differential attrition between TAP and non-TAP schools may lead to a selected sample and, for that matter, may bias the effects of the program. To investigate this possibility, we created an indicator variable that takes the value of one if the student had not ever enrolled in ninth grade in a South Carolina public school and use it as dependent variable in equation (1).[15] We utilize ninth grade enrollment for the attrition exercise because the state required students to stay in school until age 16 over the analysis period and TAP may have a direct effect on dropout over time. The estimated effect of

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

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

Table 4
Effect of TAP on long-run outcomes.

	Enrolled in 12th Grade (1)	Graduated from HS in 4 Years (2)	Any Felony (Age≤18) (3)	Any Non-Felony (Age≤18) (4)	SNAP/TANF Receipt (5)
Panel A: Difference-in-Differences Estimates					
TAP	0.035***	0.038**	−0.014**	−0.014	−0.027**
	(0.010)	(0.018)	(0.005)	(0.012)	(0.013)
Panel B: Semi-Dynamic Model Estimates					
1st Postadoption Cohort	0.021*	0.030	−0.010**	−0.012	−0.004
	(0.012)	(0.022)	(0.005)	(0.011)	(0.015)
2nd Postadoption Cohort	0.027*	0.041**	−0.015***	−0.005	−0.023
	(0.014)	(0.019)	(0.005)	(0.012)	(0.018)
3rd Postadoption Cohort	0.063***	0.073***	−0.019**	−0.036*	−0.049**
	(0.014)	(0.024)	(0.007)	(0.019)	(0.020)
4th Postadoption Cohort	0.077***	0.087**	−0.013	−0.030	−0.035*
	(0.019)	(0.039)	(0.009)	(0.020)	(0.020)
5th Postadoption Cohort	0.075***	0.086***	−0.021**	−0.048	–
	(0.022)	(0.032)	(0.010)	(0.025)	
Comparison Mean	0.673	0.664	0.045	0.164	0.607
Sample Size	43,062	38,253	43,062	43,062	30,081
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 model estimates of the effect of TAP exposure on long-run outcomes. The coefficient estimates in Panel B are obtained using imputation estimator from Borusyak et al. (2022). Standard errors are clustered at the school level. All specifications control for birth year, cohort, and school fixed effects. Student characteristics include indicators for gender, race, and free/reduced lunch status. Grade composition measures include fraction of students who are female, black, and free/reduced lunch eligible at the school-by-grade level. The dependent variable in Column 1 takes the value one if student was ever enrolled in 12th grade and it takes the value one if student graduated from high school in 4 years in Column 2. The dependent variable in Column 3 takes the value one if student was ever arrested for a felony crime as a juvenile or adult, while an analogous measure for non-felony crime is presented in Column (4). In the last column, the dependent variable takes the value one if student was ever enrolled in SNAP or TANF as an adult between ages 18 and 22. * significant at 10 %, ** significant at 5 %, *** significant at 1 %.

TAP from this analysis is 0.0004 (s.e. = 0.0041) which does not suggest any contamination due to attrition (Column 1 of Online Appendix Table A.6).

4.2. *TAP and long-run outcomes*

We present our baseline DiD results on the relationship between TAP and long-run outcomes in Panel A of Table 4. All estimates include controls for birth year, cohort, and school fixed effects, as well as student and grade composition characteristics. The DiD estimates from a specification without student and grade level controls are reported in Section 4.3. Reassuringly, the results are not sensitive to the inclusion of additional controls which provides further evidence on the quasi-randomness of the timing of TAP implementation. There are, however, efficiency gains as one can improve precision by incorporating the information contained in the observed characteristics.

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

We additionally estimate a semi-dynamic model where we allow the effect of TAP to differ depending on time-since-treatment. In doing so, we utilize the imputation estimator of (Borusyak et al., 2021) because dynamic effects, regardless of the relative period, are susceptible to bias resulting from treatment heterogeneity. The benefits of TAP may compound because students are exposed for more of their middle school years and teachers and administrators gain experience with the program. This is exactly what we find, as demonstrated in Panel B of Table 4 and visualized in Panels A and B of Fig. 1. For example, the first column indicates that the implementation of TAP increased the probability of ever being enrolled in twelfth grade by a statistically significant 2.1 percentage points for the first post-adoption cohort, while the coefficient estimate for the fourth post-adoption cohort is 7.7 percentage points. The estimated effects for on-time graduation of the same cohorts are 0.030 (s.e. = 0.022) and 0.087 (s.e. = 0.039), respectively (Column 2). Apart from highlighting the existence of a plausible dose-response relationship for educational attainment, these findings also suggest that negative weighting problem does not bias our findings. Recall that the negative weighting problem can cause DiD estimates from Panel A to fall outside the convex hull of the semi-dynamic model estimates in

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

Panel B. That is not the case here, likely due to the nontrivial size of the comparison group (Borusyak et al., 2021).

Next, we examine the effect of TAP on criminal involvement. We find that being exposed to TAP decreased the likelihood of ever being arrested for a felony crime prior to age 19 by a statistically significant 1.4 percentage points (Column 3 of Table 4). This represents a decrease of 31 percent relative to the control mean. The point estimate for non-felony offenses, reported in Column 4 of Table 4, is also negative, but smaller in magnitude (relative to the comparison mean) and indistinguishable from zero.[17] Online Appendix Table A.6 also shows the results for types of crimes, including violent crimes, alcohol- and drug-related crimes, property crimes, and other crimes, respectively (Columns 2–5). The DiD estimates for being arrested for different types of crimes at age 18 or earlier are similar in magnitude across columns.[18] For the reasons outlined in Section 3.1.2, and given the lower precision of estimates for non-felony crimes, we focus on felonies as our primary measure of justice involvement for the remainder of the paper.

We present the results on crime from the semi-dynamic specification in Panel B of Table 4. Panel C of Fig. 1 displays these results graphically. The crime-reducing effects of TAP grow over time since treatment and they are also more precisely estimated. For example, TAP adoption is associated with a 2.1 percentage point decrease in the likelihood of being arrested for a felony crime by age 18 in the fifth year of the program (Column 3). This is about twice the size of the coefficient estimate obtained in the first year of the program. As with educational attainment, this pattern is consistent with a dose-response explanation — greater exposure to TAP results in greater benefits for students.

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

To summarize the long-run results, we also construct an outcome index which is an equally weighted average of the standardized (z-scores by the academic year) measures for our key outcomes (the binary indicators for ever being arrested for a felony offense and reliance on social welfare programs are reverse coded in the construction of the index). The index allows us to obtain an estimate of the overall impact and reduces the chance of false positives due to multiple hypothesis testing (Kling et al., 2007). As shown in Column 1 of Online Appendix Table A.7, the point estimate from this exercise is 0.067 and indicates that being exposed to TAP is associated with 6.7 percent of a standard deviation increase in outcome index. Columns 2 and 3 complement this exercise by constructing alternative indices, such as one using any arrest (instead of felony arrests) and another that excludes on-time high school graduation. The estimated effect of TAP remains stable.

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 deviations higher test score value added, while the estimated impact on felony offenses maps onto attending a school with 2.4 higher standard deviations 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 SNAP or TANF 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).

We also extended our analysis to see whether there are any differential effects of TAP by gender and race. Online Appendix Table A.12 presents these results. We do not observe strong evidence of 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.

4.3. Robustness checks

We conducted several sensitivity checks to examine the robustness of our results. Since the difference-in-differences estimate nicely summarizes the effect in a single coefficient and the semi-dynamic coefficients that account for heterogeneity in impact over time align with the DiD estimates, we typically use the DiD estimate on our four key outcomes for these specification tests, displayed in Fig. 2 (details on the estimates are in Online Appendix Table A.9).

4.3.1. Alternative estimator

Online Appendix Table A.8 and Online Appendix Figure A.2 show estimates using the interaction weighted estimator outlined in Sun and Abraham (2021). The baseline findings and event studies from this alternative approach are consistent with those presented throughout the text.

4.3.2. Alternative samples

We consider the inclusion of early and late adopting cohorts (the first two robustness checks in Fig. 2). First, we exclude schools adopting TAP in the 2010–2011 academic year to see whether the estimated effects are attenuated in a meaningful way due to inflow of TAP schools as part of TIF 3 at the end of our sample period (“late adopters”). The DiD estimates are larger in magnitude. Second, our results are also robust to excluding schools adopting TAP as part of TIF 1 in the 2007–2008 academic year (“early adopters”).

We examine whether a balanced panel makes a difference in Online Appendix Figure A5, which limits the sample to cohorts from five years before and five years after TAP implementation. This restriction in the analysis sample ensures balance in event time and addresses any concerns that may arise due to the binning of pre-endpoints in Eq. (2). The pre-adoption coefficient estimates from this alternative specification closely resemble those presented in Fig. 1.

In Online Appendix Table A.10, we investigate whether differences in grade configuration across schools confound our results. Specifically, we re-estimate the baseline models by (i) excluding schools with grade configurations above ninth grade and (ii) excluding schools with grade configurations above ninth grade and those with grade configurations below fifth grade. The results are not sensitive to these sample restrictions.

4.3.3. Alternative specifications

Continuing down the estimates presented in Fig. 2, we examine the sensitivity of our results to conditioning on fifth grade composite standardized test scores. Including this control in the specifications does not

[17] Online Appendix Table A.5 presents these results by disaggregating arrests based on age (juvenile and adult). The impact of TAP continues to be larger and more precisely estimated for juvenile and adult felony offenses committed prior to age 19.

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

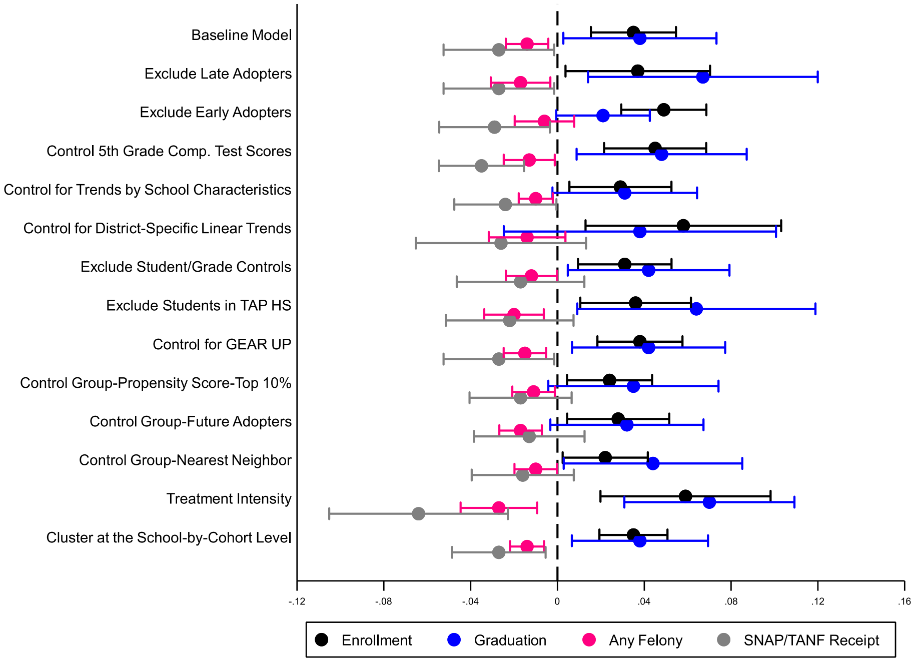


Fig. 2. Effect of TAP on long-run outcomes: robustness checks. Notes: This figure shows various robustness checks for four main outcomes. Difference-in-differences estimates for models that exclude early and late TAP adopters, control for fifth grade composite test scores, control for trends by baseline school characteristics, control for district-specific linear trends, exclude student and grade level controls, exclude students enrolled in TAP high schools, control for GEAR UP status of schools, use alternative control groups and estimate TAP effects more continuously. Each row also shows 95 % confidence interval based on standard errors clustered at the school level in Rows 1–13, while the last row displays confidence interval using standard errors clustered at the school-by-year level.

affect our estimates though the sample size is smaller due to the availability of baseline scores. We next estimate a model that interacts several baseline school-level variables 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 differ very little from those presented in Table 4. Controlling for district-specific linear pre-trends, following the two-step procedure proposed by Goodman-Bacon (2021), shows the estimated effects of TAP are similar to our main model though the confidence intervals are larger. Excluding student and grade level controls from the specifications also does not make a difference.

In the second half of Fig. 2 (“Exclude students in TAP HS”), we exclude students who ever enrolled in one of the 10 TAP high schools. The DiD estimates are similar to our baseline results.[19] 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 that South Carolina also received funding for in 2011 — is the only other education policy that may have coincided with TAP. 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.

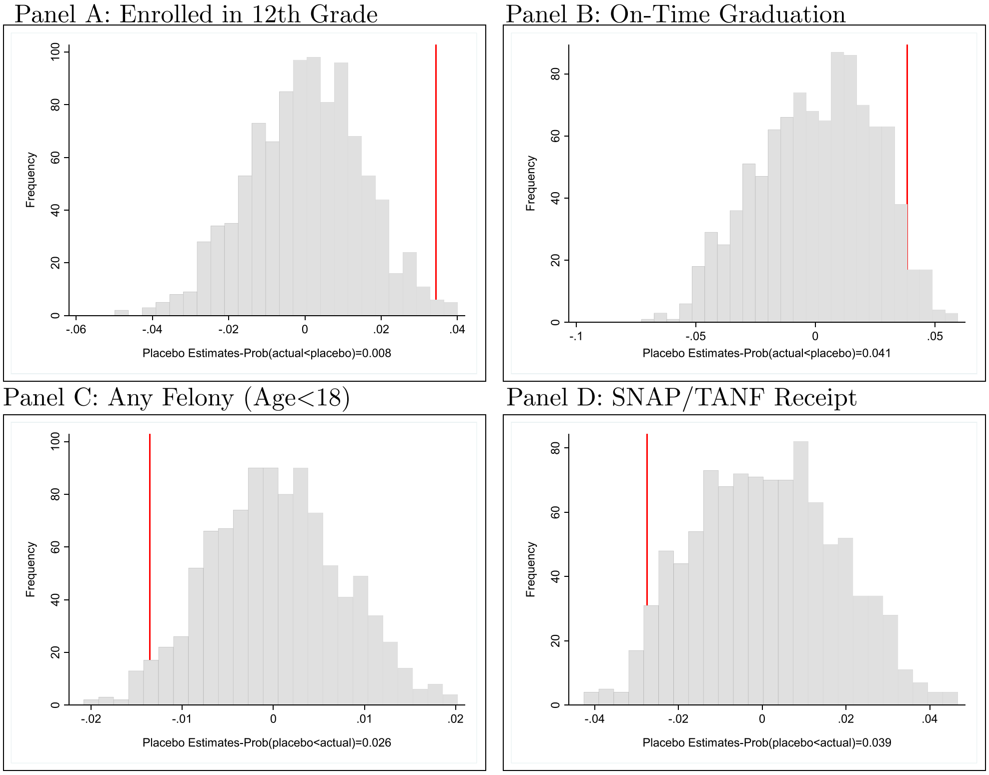
Next, our first alternative comparison group is defined by selecting the top 10 % of comparison schools based on their propensity scores (in contrast to 5 % in our main estimates). The second alternative group comprises future adopters – schools adopting TAP in 2012 (or beyond) as part of TIF 4. The point estimates using these samples are very similar to those obtained based on our primary matched control group, as are the estimates from corresponding event studies (Online Appendix Figures A3 and A4).[20] Finally, to further address the imbalance in Online Appendix Table A.2, we estimate a propensity score model by including only grade-level student characteristics. We then matched each TAP school to a comparison school using the closest propensity score. This procedure effectively eliminated differences in observable characteristics. The results from this exercise are similar to our baseline findings.

In the second to last comparison in Fig. 2, we also explore the intensity of treatment by defining TAP exposure more continuously. To do this, rather than using binary classification, we use total potential years of exposure based on the school’s grade configuration as the variable of interest. For example, total potential years of exposure for an eighth grader in the fifth year of the program in a school with grade 5–8 configuration is 4 years. The treatment in this case captures both the extensive and intensive margins. To make the results comparable to earlier results, we rescale years of exposure by dividing by the largest years of potential exposure (5 years). This ensures that treatment dosages vary between 0 and 1 and the coefficients represent the effect on the most heavily treated cohort (a change from 0 to 5). All point estimates from this alternative modeling are statistically significant and further confirm evidence of a dose-response relationship.

Our next exercises consider the structure of the standard errors. In the final row of Fig. 2, we cluster the standard errors at the school-by-year level and such alternative clustering does not affect statistical significance. In Online Appendix Table A.9, 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) to circumvent concerns over potential contamination in the inference procedure that may

[19] The DiD estimates may also in part reflect (or be confounded by) TAP exposure during elementary school years if students attended a TAP adopting elementary school. The estimated effects remain robust to excluding all the students who previously enrolled in a TAP elementary school. We present these results in Online Appendix C.

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


Panel A: Enrolled in 12th Grade

Panel B: On-Time Graduation

Panel C: Any Felony (Age<18)

Panel D: SNAP/TANF Receipt

Fig. 3. Effect of TAP on long-run outcomes: placebo estimates. Notes: This figure shows the distribution of the coefficient estimates resulting from 1000 sets of random assignments of schools to TAP adoption. The vertical lines denote the actual estimates.

arise due to small number of schools. We continue to find coefficient estimates that are statistically significant at the 5 % level.

4.3.4. Placebo tests

In addition to these sensitivity analyses, we performed two placebo tests. Our first placebo exercise shifts the analysis sample back in time to the pre-adoption period and focuses on first-time eighth graders from the 2000–2001 to 2007–2008 academic years (Online Appendix Table A.11). 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 $2003+t$.[21] As expected, we find no effects of TAP during the pre-TAP period. This indicates that the response is due to TAP and not something about the schools that implemented TAP.

Second, we randomly assign TAP adoption years to schools by drawing dates, without replacement, from the actual pool of program implementation years. We do this for 1000 sets of placebo adoptions. Fig. 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 placebo estimates that are larger (smaller) than the baseline effects on the x-axis in Panels A and B (Panels C and D). The location of the true estimates in all panels indicates that the likelihood of finding an impact merely by chance is very low.

5. Mechanisms

5.1. TAP and intermediate outcomes

We begin by presenting the impact of TAP on cognitive and non-cognitive high school outcomes in Table 5.[22] As noted earlier, there is mixed evidence on the efficacy of incentive programs on immediate academic outcomes. Thus it is important to analyze what South Carolina TAP does (or does not) do in the short-run both to understand the mechanisms behind the longer-run findings and to compare with related work.

TAP implementation influenced students well before their twelfth grade year, with the program reducing 9th grade retention and absenteeism and improving test scores. Specifically, students are 3 percentage points less likely to be retained in ninth grade (Column 1) (Online Appendix Figure A.6 presents confirmatory results from the event study specification). As shown in Column 2, on average, students in TAP schools outperformed those in comparison schools by 0.07 standard deviations on a standardized average of ELA and math scores from the 10th grade exit exam. Finally, TAP led to fewer days of absence in tenth grade, though this difference is not statistically significant. These school outcomes have different availability than the longer-term outcomes, so we check for robustness to attrition in Online Appendix Section A.1.1, finding little scope for attrition to undermine the findings. We also consider whether exposure to TAP changed the high schools students attended, finding little difference (Online Appendix Section A.1.2). In short, exposure to TAP improved students' performance throughout their high school trajectory.

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

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

[22] The available data for these intermediate outcomes vary. South Carolina had an exit exam from 2006 to 2015, the High School Assessment Program (HSAP) which consisted of English language arts and math exams administered in the spring of 10th grade. 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.

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)	0.069** (0.029)	−2.364 (2.185)
Comparison Mean	0.120	−0.102	18.22
Sample Size	40,800	27,096	26,323
Controls:			
Cohort and School Fixed Effects	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes
Grade Composition (8th Grade)	Yes	Yes	Yes

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

specification of the following form:

$$IO^j_{isc} = \alpha^j_1 TAP_{sc} + X'_{isc}\alpha_2 + \delta_s + \lambda_c + \epsilon_{isc} \tag{3}$$

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

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

where β^{Res}_{DiD} captures the component of the estimated TAP effect that is not explained by improvements in intermediate outcomes which also can be expressed as $\beta_{DiD} = \beta^{Res}_{DiD} + \sum_j \alpha^j_1 \theta^j$. 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 compute its explanatory power by $\frac{\alpha^j_1 \theta^j}{\beta_{DiD}}$ and display the results in Online Appendix Table A.13 and Online Figure A.7. The estimates in this exercise are limited to students with non-missing test scores and non-test outcomes. The stability of the DiD coefficient with the limited sample provides assurance that our results are not confounded by the introduction of exit exams in South Carolina public schools.

For example, the results from this exercise imply that the decrease in the probability of being retained in ninth grade explains approximately 15 percent of the on-time graduation treatment effect, tenth grade composite test scores explain up to 14 percent of the same effect and student absenteeism, which is a proxy for non-cognitive ability (Gershenson, 2016; Holbein and Ladd, 2017; Jackson, 2018; Jackson et al., 2020), explains 16 percent of the on-time graduation effect. Averaging across the four main outcome variables, the explained portion of the estimated effect of TAP is around 47 percent, with greater explanatory power for school outcomes (enrollment and graduation) and less for crime and welfare outcomes.[23]

5.2. *TAP, teachers and school culture*

Although the predictive power of high school outcomes is non-trivial, our results do not speak to the question of *why* we observe favorable intermediate outcomes for students in TAP schools. To further understand how TAP improved high school outcomes, we consider the following school-level domains which are known to be associated with improvements in student well-being: the teacher workforce in 8th grade and school climate. Table 6 presents evidence on whether TAP led to changes in the composition of the teacher workforce. The total number of teachers in TAP schools remained constant; however, the program increased turnover by around 4 percent relative to the comparison mean (Columns 1 and 2). Considering the average total number of teachers in comparison schools approximately 32, this increase in turnover is roughly equivalent to a new hire in TAP schools per year in the post-adoption period.

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

The reduction in teacher qualifications and experience 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 that information asymmetries increase with tenure, inter-school mobility is more prevalent among experienced teachers post-awards (Bates, 2020; Berlinski and Ramos, 2020).

This minor increase in turnover and resultant changes in the teacher workforce are unlikely to generate the long-run effects observed throughout the paper. For one, these teacher changes would generally be detrimental to student outcomes—the opposite of our findings. To shed further light on the role of teacher sorting in explaining our findings, we reestimate our results in Online Appendix Table A.14 by excluding

[23] We also explore whether time spent in school drives the reduction in crime, finding little evidence for an explanation related to incapacitation effect of schooling (Online Appendix Section A.1.3).

Table 6
Mechanisms: effect of TAP on 8th grade school characteristics.

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

Notes: This table reports difference-in-differences estimates of the effect of TAP exposure on 8th grade school characteristics. Highly qualified status of a teacher is defined using a combination of state certification and subject matter knowledge scores in state assessments. All specifications control for fraction of students who are female, black and free/reduced lunch eligible at the school-by-year level. Standard errors are clustered at the school level. * significant at 10 %, **significant at 5 %.

Table 7
Mechanisms: effect of TAP on teacher, student and parent satisfaction.

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

Notes: This table reports difference-in-differences estimates of the effect of TAP exposure on school climate surveys. The overall satisfaction index, reported in the first column, includes the percentage 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 as well as total school enrollment. Standard errors are clustered at the school level. N represents the sample size.

TAP schools with high teacher turnover rates.[24] As shown in the first column, the impact of TAP on teacher turnover is almost equal to zero in magnitude for this selected sample, while the long-run improvements in student outcomes due to TAP on student outcomes are consistently larger than those reported in Table 4 and continue to be statistically significant. We also show in Online Appendix Section A.1.4 that there is little scope for changes in principals to explain the TAP effects.

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 estimate TAP impacts on responses to a survey administered annually to teachers, students, and parents which is an integral part of the state's accountability system (Table 7). 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." We also create an overall index of satisfaction, which is defined separately for each group of respondents, by averaging the z-scores of satisfaction measures.

As shown in Panel A of Table 7, being exposed to TAP is associated with 0.42 standard deviation increase in the parental satisfaction index (Column 1). The coefficient estimates for the individual components (Columns 2–4) suggest improvements in all domains. Interestingly, the impact of TAP on student satisfaction 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 as well as the social and physical environment increased following the adoption of TAP, but the coefficient estimates fall short of statistical significance. Overall, these results align well with explanations related to changes in school climate as well as increases in the productivity of incumbent teachers.

[24] These are TAP schools that experienced an average change in teacher turnover rates of more than 5 percentage points (in absolute value) between the pre- and post-adoption periods.

5.3. Contributions of the program components

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

A growing body of research casts doubt on the efficacy of teacher professional development programs 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 program components that parallel traditional professional development — multiple career paths and ongoing applied professional training — to account for the entire impact of TAP on student outcomes. However, there is evidence from other settings that teacher observations and feedback can improve student performance (Taylor and Tyler, 2012; Briole and Maurin, 2022; Taylor, 2023) which aligns with the TAP element of instruction-focused accountability.[25]

In an attempt to shed light on the role and design of incentives, we compare South Carolina TAP with its adoption in the Chicago Public Schools via a randomized controlled trial of 34 high-need schools (K-8) from 2007 to 2010. South Carolina and Chicago TAP were identical in design. However, the individual teacher value-added component of performance-based compensation was not implemented in Chicago TAP because the data needed to reliably link students and teachers were not available. Therefore, the Chicago program used group incentives (i.e., school-level value-added) to calculate performance bonuses for teachers (Glazerman and Seifullah, 2012). Other aspects of the incentive pay were quite similar across locations: mentor and master teachers received comparable compensation for their duties and the average annual bonus pay for teachers was around $2000 in Chicago TAP. This latter amount represents a 3.3 percent change in base teacher salary in Chicago, while the average bonus pay is approximately 4.3 percent of the base salary in South Carolina. The student bodies in both intervention sites were also similarly high-need students, and while the grade configurations of treated schools were different, we show elsewhere that our results are not sensitive to grade configuration (Online Appendix Table A.10). Thus, the most notable distinction between the TAP programs is the existence of test-based individual teacher performance pay in South Carolina TAP in concert with group-based incentives, whereas Chicago solely provided group-based incentives.

In their evaluation of the Chicago program, Glazerman and Seifullah (2012) find that Chicago TAP did *not* lead to improvements in student achievement, nor did the effects grow with longer years of exposure. This lack of improvement holds true for both lower (grades 4–6) and upper grades (grades 7–8) individually throughout the entire program, contrasting with our South Carolina evidence showing benefits on short-run academic outcomes (i.e., lower ninth grade retention and higher exit exam scores). Similar short-run program benefits were also found at other TAP sites which fully implemented the program including individual incentives (Springer et al., 2014; Chiang et al., 2015; Eren, 2019).

While we recognize that the discrepancy in these results may not be solely attributable to (unintended) omission of individual incentives, this comparison implies that tying compensation to teacher value-added scores in South Carolina TAP was a crucial component of program success. However we note that individual incentives on their own may not be sufficient for improving student outcomes. Existing evidence regarding the impact of a stand-alone pay scheme focusing exclusively on test-based individual incentives in the U.S. (such as the program studied in Springer et al. (2010)) *also* does not boost student outcomes. Thus we conclude that TAP's effectiveness may come from complementarities within the program itself: individual incentives are necessary but not sufficient.

6. Benefit-cost analysis of TAP

In this section, we provide a simple back-of-the envelope cost calculation to put these estimated impacts into monetary perspective. Before proceeding, it is important to keep in mind that any benefit-cost analysis is speculative and subject to several caveats. The total average cost of TAP implementation is about $250 per student (Institute of Education Sciences, 2015). We break the benefits associated with TAP into two components: (i) broader benefits to society originating from reduced crime and (ii) future gains due to increased high school graduation. Recent research suggests that receipt of government assistance — a cost to taxpayers — leads to a wide range of positive outcomes, including improved adult health, better birth and child outcomes and lower criminal involvement (Almond et al., 2011; Hoynes et al., 2016; Tuttle, 2019). Because of this uncertainty in net social gains, we opt out of including the benefit to taxpayers resulting from reduced reliance on SNAP/TANF 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 are based on jury award data and we use per victim cost values. For each individual in our analysis sample, we obtain an overall social cost of crime by summing victim cost values from all arrests up to age 18.[26] We use this total cost measure as variable of interest in equation (1) and estimate the impact of TAP on social benefits resulting from averted crimes.[27] Panel A of Table 8 reports the point estimates from this exercise for any criminal activity and felony offenses in rows 1 and 2, respectively. The estimated benefits from reduced crime outweigh the cost of TAP by more than 6 to 1.

Next, we follow Heller et al. (2017) in our calculations of future monetary gains due to increased high school graduation. We assume that each graduate accrues one additional year of education relative to each non-graduate and focus on values related to earnings and health. To estimate the gains associated with earnings, we use synthetic work-life estimates from Julian and Kominski (2011). Work-life earnings represent 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

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

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

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

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 are 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 %.

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 is $346, which we proxy by expenditures per pupil from the Common Core of Data averaged over 2006–2016 ($9,932). Our estimate for the net future benefits of graduation from Panel B is around $2037. Combining the benefits from reduced felony offenses and increased graduation results in an MVPF of 14, making TAP a very cost-effective program.

7. Conclusion

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

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

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered potential competing interests:

Given her role as a co-editor Journal of Public Economics, Sarah R. Cohodes had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to another journal editor. [SRC] If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.jpubeco.2025.105561.

Data availability

The data that has been used is confidential.

References

Abdulkadiroğlu, A., Angrist, J.D., Hull, P.D., Pathak, P.A., 2016. Charters without lotteries: testing takeovers in New Orleans and Boston. Am. Econ. Rev. 106 (7), 1878–1920.
Abeler, J., Huffman, D.B., Raymond, C., 2023. Incentive Complexity, Bounded Rationality and Effort Provision. Technical report, IZA Discussion Papers.
Abeler, J., Jäger, S., Aug 2015. Complex tax incentives. Am. Econ. J.: Econ. Policy 7 (3), 1–28.
Agan, A., Garin, A., Koustas, D., Mas, A., Yang, C.S., 2024. The labor market impacts of reducing felony convictions. Am. Econ. Rev.: Insights 6, 341–358.
Aizer, A., Doyle, J.J., 2015. Juvenile incarceration, human capital, and future crime: evidence from randomly assigned judges. Q. J. Econ. 130, 759–803.
Almond, D., Hoynes, H.W., Schanzenbach, D.W., 2011. Inside the war on poverty: the impact of food stamps on birth outcomes. Rev. Econ. Stat. 93 (2), 387–403.
Alsan, M., Barnett, A.H., Hull, P., Yang, C., 2024. Something works in U.S. Jails: misconduct and recidivism effects of the ignite program. NBER Working Paper: 32282.
Anders, J., Barr, A.C., Smith, A.A., Feb 2023. The effect of early childhood education on adult criminality: evidence from the 1960s through 1990s. Am. Econ. J.: Econ. Policy 15 (1), 37–69.
Angrist, J.D., Guryan, J., 2008. Does teacher testing raise teacher quality? Evidence from state certification requirements. Econ. Educ. Rev. 27 (5), 483–503.
Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H., Wilson, D., 2009. Evaluating the impact of performance-related pay for teachers in England. Labour Economics 16 (3), 251–261.
Bailey, D., Duncan, G.J., Odgers, C.L., Yu, W., 2017. Persistence and fadeout in the impacts of child and adolescent interventions. J. Res. Educ. Eff. 10 (1), 7–39.
Bailey, D.H., Duncan, G.J., Cunha, F., Foorman, B.R., Yeager, D.S., 2020. Persistence and fade-out of educational-intervention effects: mechanisms and potential solutions. Psychol. Sci. Public Interest 21 (2), 55–97.
Bates, M., 2020. Public and private employer learning: evidence from the adoption of teacher value added. J. Labor Econ. 38 (2), 375–420.
Berlinski, S., Ramos, A., 2020. Teacher mobility and merit pay: evidence from a voluntary public award program. J. Public Econ. 186, 104186.
Beuermann, D.W., Jackson, C.K., 2022. The short-and long-run effects of attending the schools that parents prefer. J. Hum. Resour. 57 (3), 725–746.
Biasi, B., Aug 2021. The labor market for teachers under different pay schemes. Am. Econ. J.: Econ. Policy 13 (3), 63–102.
Bonilla, S., Dee, T.S., 2020. The effects of school reform under Nclb waivers: evidence from Focus Schools in Kentucky. Educ. Finance Policy 15 (1), 75–103.
Borman, G.D., Hewes, G.M., Overman, L.T., Brown, S., 2003. Comprehensive school reform and achievement: a meta-analysis. Rev. Educ. Res. 73 (2), 125–230.
Borman, G.D., Slavin, R.E., Cheung, A.C.K., Chamberlain, A.M., Madden, N.A., Chambers, B., 2007. Final reading outcomes of the national randomized field trial of success for all. Am. Educ. Res. J. 44 (3), 701–731.
Borusyak, K., Jaravel, X., Spiess, J., Revisiting event study designs: Robust and efficient estimation, arXiv preprint arXiv:2108.12419, 2021.

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

Briole, S., Maurin, E., 2022. There's always room for improvement: the persistent benefits of a large-scale teacher evaluation system. J. Hum. Resour. 1220–11370.

Callaway, B., Sant'Anna, P.H.C., 2020. Difference-in-differences with multiple time periods. J. Econ..

Cameron, A.C., Gelbach, J.B., Miller, D.L., 2008. Bootstrap-based improvements for inference with clustered errors. Rev. Econ. Stat. 90 (3), 414–427.

Carneiro, P., Cruz-Aguayo, Y., Intriago, R., Ponce, J., Schady, N., Schodt, S., 2022. When promising interventions fail: personalized coaching for teachers in a middle-income country. J. Public Econ. Plus 3, 100012.

Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D.W., Yagan, D., 2011. How does your kindergarten classroom affect your earnings? Evidence from project STAR. Q. J. Econ. 126 (4), 1593–1660.

Chiang, H., Wellington, A., Hallgren, K., Speroni, C., Herrmann, M., Glazerman, S., Constantine, J., 2015. Evaluation of the Teacher Incentive fund: implementation and impacts of pay-for-performance after two years. Ncee 2015-4020. Natl. Cent. for Educ. Eval. and Reg. Assist..

Choi, J.J., Laibson, D., Madrian, B.C., 2009. Reducing the complexity costs of 401 (k) participation through quick enrollment. In: Developments in the Economics of Aging. University of Chicago Press, pp. 57–82.

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

Cohodes, S.R., 2020. The long-run impacts of specialized programming for high-achieving students. Am. Econ. J.: Econ. Policy 12 (1), 127–166.

Cohodes, S.R., Grossman, D.S., Kleiner, S.A., Lovenheim, M.F., 2016. The effect of child health insurance access on schooling: evidence from public insurance expansions. J. Hum. Resour. 51 (3), 727–759.

Cook, P.J., Kang, S., 2016. Birthdays, schooling, and crime: regression-discontinuity analysis of school performance, delinquency, dropout, and crime initiation. Am. Econ. J.: Appl. Econ. 8 (1), 33–57.

Currie, J., Grogger, J., Burtless, G., Schoeni, R.F., 2001. Explaining recent declines in food stamp program participation. Brook.-Whart. Pap. Urban Aff. 203–244.

Cutler, D.M., Lleras-Muney, A., Jul 2006. Education and Health: Evaluating Theories and Evidence. Working Paper 12352, National Bureau of Economic Research.

Davis, J.M.V., Heller, S.B., 2020. Rethinking the benefits of youth employment programs: the heterogeneous effects of summer jobs. Rev. Econ. Stat. 102 (4), 664–677.

Dee, T.S., Wyckoff, J., 2015. Incentives, selection, and teacher performance: evidence from impact. J. Policy Anal. Manag. 34 (2), 267–297.

Deming, D.J., 2011. Better schools, less crime? Q. J. Econ. 126 (4), 2063–2115.

Deming, D.J., Hastings, J.S., Kane, T.J., Staiger, D.O., 2014. School choice, school quality, and postsecondary attainment. Am. Econ. Rev. 104 (3), 991–1013.

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

Dynarski, S.M., Scott-Clayton, J.E., 2006. The cost of complexity in federal student aid: lessons from optimal tax theory and behavioral economics. Natl. Tax J. 59 (2), 319–356.

Englmaier, F., Roider, A., Sunde, U., 2017. The role of communication of performance schemes: evidence from a field experiment. Manag. Sci. 63 (12), 4061–4080.

Eren, O., 2019. Teacher incentives and student achievement: evidence from an advancement program. J. Policy Anal. Manag. 38 (4), 867–890.

Figlio, D., Kenny, L., Jun 2007. Individual teacher incentives and student performance. J. Public Econ. 91 (5–6), 901–914.

Fryer, R.G., 2013. Teacher incentives and student achievement: evidence from New York City public schools. J. Labor Econ. 31 (2), 373–407.

Fryer, R.G., 2014. Injecting charter school best practices into traditional public schools: evidence from field experiments. Q. J. Econ. 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, vol. 2. Elsevier, pp. 95–322.

Garet, M.S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K., Falk, A., Bloom, H.S., Doolittle, F., 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? J. Inst. Theor. Econ./Z. ges. Staatswiss. 200–221.

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

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

Glewwe, P., Ilias, N., Kremer, M., Jul 2010. Teacher incentives. Am. Econ. J.: Appl. Econ. 2 (3), 205–227.

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

Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. J. Econ.

Gray-Lobe, G., Pathak, P.A., Walters, C.R., 2021. The Long-Term Effects of Universal Preschool in Boston. Technical report, National Bureau of Economic Research.

Hanushek, E.A., Luo, J., Morgan, A.J., Nguyen, M., Ost, B., Rivkin, S.G., Shakeel, A., Mar 2023. The Effects of Comprehensive Educator Evaluation and Pay Reform on Achievement. Working Paper 31073, National Bureau of Economic Research.

Heckman, J., Pinto, R., Savelyev, P., 2013. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. Am. Econ. Rev. 103 (6), 2052–2086.

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

Hendren, N., Sprung-Keyser, B., 2020. A unified welfare analysis of government policies. Q. J. Econ. 135 (3), 1209–1318.

Hill, A.J., Jones, D.B., 2020. The impacts of performance pay on teacher effectiveness and retention: does teacher gender matter? J. Hum. Resour. 55 (1), 349–385.

Hjalmarsson, R., Holmlund, H., Lindquist, M.J., 2015. The effect of education on criminal convictions and incarceration: causal evidence from micro-data. Econ. J. 125 (587), 1290–1326.

Holbein, J.B., Ladd, H.F., 2017. Accountability pressure: regression discontinuity estimates of how no child left behind influenced student behavior. Econ. Educ. Rev. 58, 55–67.

Holmstrom, B., 1982. Moral hazard in teams. Bell J. Econ. 13 (2), 324–340.

Holmstrom, B., Milgrom, P., 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. J. L. Econ. & Org. 7, 24.

Hoynes, H., Schanzenbach, D.W., Almond, D., 2016. Long-run impacts of childhood access to the safety net. Am. Econ. Rev. 106 (4), 903–934.

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

Imberman, S.A., Lovenheim, M.F., 2015. Incentive strength and teacher productivity: evidence from a group-based teacher incentive pay system. Rev. Econ. Stat. 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. J. Polit. Econ. 126 (5), 2072–2107.

Jackson, C.K., Bruegmann, E., 2009. Teaching students and teaching each other: the importance of peer learning for teachers. Am. Econ. J.: Appl. Econ. 1 (4), 85–108.

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

Julian, T., Kominski, R., 2011. Education and synthetic work-life earnings estimates. American Community Survey reports. Acs-14. US Census Bureau.

Kane, T.J., Staiger, D.O., Dec 2008. Estimating Teacher Impacts on Student Achievement: an Experimental Evaluation. Working Paper 14607, National Bureau of Economic Research.

Kleven, H.J., Kopczuk, W., 2011. Transfer program complexity and the take-up of social benefits. Am. Econ. J.: Econ. Policy 3 (1), 54–90.

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

Kling, J.R., Ludwig, J., Katz, L.F., 2005. Neighborhood effects on crime for female and male youth: evidence from a randomized housing voucher experiment. Q. J. Econ. 120 (1), 87–130.

Lavy, V., 2002. Evaluating the effect of teachers' group performance incentives on pupil achievement. J. Polit. Econ. 110 (6), 1286–1317.

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. Rev. Econ. Stud. 87 (5), 2322–2355.

Levitt, S.D., Lochner, L., 2001. The determinants of juvenile crime. In J. Gruber (Ed.), Risky Behavior among Youths: An Economic Analysis 7, 327–373.

Lochner, L., Moretti, E., Mar 2004. The effect of education on crime: evidence from prison inmates, arrests, and self-reports. Am. Econ. Rev. 94 (1), 155–189.

Loyalka, P., Popova, A., Li, G., Shi, Z., 2019. Does teacher training actually work? Evidence from a large-scale randomized evaluation of a national teacher training program. Am. Econ. J.: Appl. Econ. 11 (3), 128–154.

Ludwig, J., Miller, D.L., 2007. Does head start improve children's life chances? Evidence from a regression discontinuity design. Q. J. Econ. 122 (1), 159–208.

Miller, T.R., 1996. Victim Costs and Consequences: A New Look. US Department of Justice, Office of Justice Programs.

Morgan, A.J., Nguyen, M., Hanushek, E.A., Ost, B., Rivkin, S.G., Mar 2023. Attracting and Retaining Highly Effective Educators in Hard-to-Staff Schools. Working Paper 31051, National Bureau of Economic Research.

Muralidharan, K., Sundararaman, V., 2011. Teacher performance pay: experimental evidence from India. J. Polit. Econ. 119 (1), 39–77.

Rockoff, J.E., Jacob, B.A., Kane, T.J., Staiger, D.O., Jan 2011. Can you recognize an effective teacher when you recruit one? Educ. Finance Policy 6 (1), 43–74.

Rose, E.K., Schellenberg, J.T., Shem-Tov, Y., 2022. The Effects of Teacher Quality on Adult Criminal Justice Contact. Technical report, National Bureau of Economic Research.

Schueler, B.E., Asher, C.A., Larned, K.E., Mehrotra, S., Pollard, C., 2022. Improving low-performing schools: a meta-analysis of impact evaluation studies. Am. Educ. Res. J. 59 (5), 975–1010.

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

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

South Carolina Department of Education, 2012. The System for Teacher and Student Advancement.

Springer, M.G., Ballou, D., Peng, A., 2014. Estimated effect of the teacher advancement program on student test score gains. Educ. Finance Policy 9 (2), 193–230.

Springer, M.G., Hamilton, L., McCaffrey, D.F., Ballou, D., Le, V.-N., Pepper, M., Lockwood, J.R., Stecher, B.M., 2010. Teacher pay for performance: experimental evidence from the project on incentives in teaching. Natl. Cent. on Perform. Incent.

Staiger, D.O., Rockoff, J.E., Sep 2010. Searching for effective teachers with imperfect information. J. Econ. Perspect. 24 (3), 97–118.

Sun, L., Abraham, S., 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J. Econ. 225 (2), 175–199.

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

Taylor, E.S., Tyler, J.H., Dec 2012. The effect of evaluation on teacher performance. Am. Econ. Rev. 102 (7), 3628–3651.

Tuttle, C., 2019. Snapping back: food stamp bans and criminal recidivism. Am. Econ. J.: Econ. Policy 11 (2), 301–327.

Zimmer, R., Henry, G.T., Kho, A., 2017. The effects of school turnaround in tennessee's achievement school district and innovation zones. Educ. Eval. Policy Anal. 39 (4), 670–696.