Scaling Up: Advanced Placement Incentive Program

Hande Celebi
University of Texas at Austin
handenurcelebi@utexas.edu
for the updated version click here.

December 2023

Abstract

This paper explores the success of scaling up of state programs. I study the two phases of staggered rollout of the Advanced Placement (AP) Incentives Programs in Texas which aimed to increase AP utilization. The 1997 pilot phase included a small group of high schools, the Scaled-up phase expanded this to half of all Texas high schools in 2001. Using a staggered difference-in-differences approach, I separately estimate the impact of the program in each phase. I find a 70% increase in AP Enrollment and a 40% increase in the number of AP courses for the schools which implemented the program before scaling up. College enrollment and graduation increases by 10% and 3%, as well as a 7% increase in wages. However there is a null effect for the outcomes of the schools which implemented the program after scaling up. the potential explanation for the disparate effects is that the student demand for the AP courses differed between two treatment arms.

Keywords: Scaling up, Advanced Placement, Callaway&Sant'Anna, Variation in Treatment Timing, High School Education
JEL classification: I26, I28

1 Introduction

Since its inception in 1951, Advanced Placement (AP) courses have provided students with an opportunity to excel during their high school years, demonstrate their abilities to college admission committees, and earn college credits before entering college. Enhancing student achievement has consistently been a core objective of federal and state-level education policies. The prevalence of AP participation has steadily increased over time, with approximately 71% of public schools in the U.S. incorporating AP programs into their curriculum by 2009 (Theokas and Saaris n.d.). In certain states like Arkansas, Indiana, and

Mississippi, the inclusion of AP programs has become mandatory (Klopfenstein and Thomas 2009). Consequently, numerous policies have been designed over time to provide incentives for offering and taking AP courses.

The Advanced Placement Incentive Program (APIP) is one such incentive program designed to recognize and reward AP students, teachers, and schools. APIP was introduced in 1993 in the state of Texas and provides funding for teachers who teach AP courses, students who take these courses, and schools that offer them. The program initially started on a small scale, with funding obtained from private donors, and schools had to wait until the program found a donor for them after applying. Over time, the total funding increased with the assistance of federal policies like No Child Left Behind in 2001, and the financial constraint on accessing APIP funds was lifted.

In the field of applied research, natural experiments or randomized control trials are employed to establish a causal relationship between policies and their intended outcomes. Researchers and policymakers often implement policy changes in small areas like cities or neighborhoods and analyze the outcomes in these specific locales. If positive results are observed, the new policy may be extended to a larger area, with the expectation of similar improvements in outcomes. However, recent literature has provided examples where scaling up a program did not lead to the expected positive outcomes, raising concerns about the scalability of experimental results (A. Banerjee et al. 2017, Al-Ubaydli, List, and Suskind 2017, Duflo 2004).

This paper investigates whether APIP was effective both before and after scaling up in achieving its goal of increasing AP utilization in the state of Texas. Additionally, the study evaluates the long-term impact of APIP beyond its initial objectives. Moreover, the seeting allows to investigate the reasons and processes behind the failure of a scaling up plan by examining the effects of the Advanced Placement Incentive Program (APIP) before and after it was scaled up.To achieve this, the study leverages differences in the timing of high schools receiving APIP funding for the first time.

To estimate the effects, a difference-in-differences (DiD) model is employed using the Callaway & Sant'Anna method (Callaway and Sant'Anna 2021) to recover the average treatment effect on the treated (ATT) for school-level outcomes and the intention-to-treat effect (ITT) for individual-level outcomes. Using a generalized propensity score as the first step of Callaway & Sant'Anna method, the timing of staggered APIP implementation is aimed to be exogenous to the conditions and characteristics of the schools before the treatment. The sample is split into two parts: schools treated before 2001 and schools treated between 2001 and 2005. The analysis is conducted as if two different policies were implemented for these groups of schools. The control schools for each treatment arm are those that implemented APIP after 2005 or never implemented it.

I focus on all secondary schools and their students in Texas, utilizing the Texas Education Research Center (ERC) dataset. This unique longitudinal dataset covers a 10-year period, including all students who enrolled in a secondary school in Texas from 1995 to 2005 and their long term outcomes of college and employment status from 1995 to 2020. It contains valuable information on student demographics, courses taken, pass/fail status, as well as school-related data such as funding, teacher demographics, and wages. Furthermore, the dataset allows the tracking of students for at least 15 years after their secondary school graduation, enabling the examination of long-term effects. Key outcome variables, namely AP enrollment, number of AP courses, count of college enrollees, college graduates, employment rate, and mean wages are synthesized at the school level by aggregating student numbers across academic years.

The sample is divided into pilot schools (those that implemented the program between 1997-2000) and scaled-up schools (those that implemented the program between 2001-2005). The comparison group consists of schools that either implemented the program after 2005 or never implemented it. The study's findings reveal that for pilot schools, APIP leads to a 70% increase in AP enrollment (from 163 students to 247 students) and a 40% increase in the number of AP courses offered (from 4.5 courses to 6.5). However, there are no statistically or economically significant changes observed in scaled-up schools. pilot schools experience positive long-run outcomes, including a 24% increase in college enrollment and a 6% increase in college graduation rates, as well a 7% increase in wages at the ages between 25-30 conditional on high school graduation. In contrast, all of these outcomes are not statistically or economically significant for scaled-up schools.

To comprehend the underlying mechanism behind the differences between pilot and scaled-up schools, several potential reasons are considered. Firstly, geographical factors such as location in cities or proximity to each other could contribute to the success of pilot schools. Secondly, highly motivated schools might be more likely to apply for the APIP funding earlier. Third, APIP being a small program run by a local agency might increase the probability of being observed, so pilot schools might work more towards APIP goals due to the Hawthorne Effect. Forth, differences in supply-side mechanisms, such as offering more qualified AP teachers could influence the outcomes. Fifth, differences in demand-side mechanisms, such as experiencing higher demand for AP courses, could also influence the outcomes. The study finds evidence consistent with a demand mechanism, as pilot schools already had higher AP class sizes, and this trend persisted after the introduction of new AP courses through APIP. Moreover, scaled-up schools with higher AP class sizes are able to increase AP enrollment. Larger class sizes can be seen as a signal of high demand for AP courses in pilot schools.

The outcomes raise a discussion on the inclusivity of the AP programs. Although the scaled-up schools do not increase AP utilization, there is a possiblity that those schools make AP program more inclusive for disadvantaged students. There are studies presenting evidence that AP programs are not able to include disadvantaged schools or students (Klopfenstein 2004, Long, Conger, and McGhee Jr 2019, Hallett and Venegas 2011). Since one of the incentives of the APIP is a means-tested test fee reduction, one of the targets of the program should be low-income students. As evidenced by the higher increase in higher income schools and no change in the composition of AP students, I can conclude that APIP is not successfully include disadvantaged students.

The paper makes three significant contributions. Firstly, it extends the existing literature on scaling up by presenting a unique example of a natural experiment involving a large-scale educational program encompassing all students. Numerous instances in education economics literature demonstrate unsuccessful scale-ups. Some studies utilize randomized control trials, or field experiments in a small scale, find positive results (Duflo, Dupas, and Kremer 2011, A. V. Banerjee et al. 2007, Fryer, Levitt, List, et al. 2015), but replications in a slightly different settings (Bold et al. 2012), in another country (Barrera-Osorio and Linden 2009), or in a larger area (Andrabi et al. 2020, List 2022) end with a failure. These papers echoes the findings of Banerjee et al. (A. Banerjee et al. 2017) who identified six main reasons for the failure of scaling up: market equilibrium effects, spillovers, political reactions, context dependence, randomization or site-selection bias, and piloting bias.

There are also studies outside of education economics that show unsuccessful scale-ups. Mobarak (Mobarak 2022) points out that the pilot study they conduct in 2008 to migrate seasonal workers to the cities for nutritional purposes in Bangladesh was successful and they scaled the program in stages. Scaling up , however, resulted in higher divorce rates, higher prices in cities and higher family separation. Many papers in health economics point out that some treatments do not work at scale because of a decrease in the quality of service, lack of human resources, high cost or different political motives (Monroe-DeVita, Morse, and Bond 2012; Kurowski et al. 2007; Kumaranayake 2008; Bold et al. 2012).

Secondly, the paper sheds light on the significance of providing access to Advanced Placement (AP) courses, as evidenced by increased college enrollment and graduation rates due to AP utilization. The literature consistently demonstrates positive outcomes associated with taking AP courses. Despite using a limited portion of all Texas districts and being unable to utilize staggered program implementation, Jackson (Jackson 2010) highlighted that APIP implementation was associated with higher AP exam participation, as well as higher SAT scores and college matriculation although he was not able to show a positive effect of APIP on AP enrollment. By applying a administrative dataset from Texas and new methods as Callaway & Sant'Anna difference-in-differences, I am able to investigate long run effects and improve the estimation quality. Smith et al. (Smith, Hurwitz, and Avery 2017) utilized a novel dataset with observable AP examination scores, finding that higher AP examination scores positively affect college completion and subsequent exam-taking, employing a regression discontinuity design. Additionally, several studies establish a positive correlation between AP participation and academic performance, including college attendance (Chajewski, Mattern, and Shaw 2011) and college completion (Hargrove, Godin, and Dodd 2008; Mattern, Marini, and Shaw 2013).

Thirdly, the paper contributes to the literature analyzing the association between additional school resources and student achievement by highlighting the importance of assessing the demand of a program contributes so much to the success of a financial resource program. This paper presents that an increase of

$1.32 per student increases the 4-year college graduation by 6% for the schools which implement the program before scaling up. However, the overall effect of the program on college graduation is not statistically different than zero. Recent literature utilizing rigorous causal inference methods has estimated a positive causal effect of higher teacher salaries (Ferguson 1991), class size (Krueger 1999), capital infrastructure (Hong and Zimmer 2016), and expenditure per-pupil (Gigliotti and Sorensen 2018) on student achievement. A meta-analysis on design-based studies, which are plausibly causal, present evidence that a policy increasing spending by $1000 per-pupil for four years improves test scores by 0.0316sd and college-going by 2.8pp (Jackson and Mackevicius n.d.).

The identification of these effects is of paramount significance for state and local education agencies, as it facilitates optimal resource allocation to maximize educational impact. The indiscriminate distribution of APIP funds to districts struggling to bolster AP participation due to multifaceted reasons underscores inefficiency in resource utilization. Pinpointing the challenges faced by these districts and educational institutions in augmenting AP participation can enable targeted investments tailored to address specific impediments.

Furthermore, a deepened understanding of the circumstances under which the scaling up of a policy may not yield anticipated outcomes, thus culminating in the squandering of valuable resources, is of utmost importance. A successful scale-up demands meticulous consideration of variables such as false positives, the representativeness of initial scenarios or populations, the potential for spillover effects, and the costs that might become prohibitive. Our analysis aligns with this imperative, mirroring findings from a plethora of studies within the realm of education economics and health economics. These studies often unravel instances where programs that exhibited promise in small-scale trials faltered when expanded to larger contexts.

The structure of the paper is as follows: Section 2 explains the AP courses and APIP, Section 3 describes the data, Section 4 presents the empirical methods used in the estimations and their internal validity, Section 5 presents the results, robustness checks, and several heterogeneity analyses, Section 6 explores possible mechanisms behind the results, Section 7 discusses diversity and cost-benefit of the APIP, and Section 8 concludes.

2 Setting

2.1 Advanced Placement Program

Advanced Placement (AP) courses are designed for secondary school students to gain exposure to college-level academic content during their high school years. The primary objectives of these courses are to prepare students for advanced academic challenges and post-secondary education. By enrolling in AP courses, students can enhance their intellectual abilities, demonstrate their potential to college admission committees, and earn college credits while still in secondary school. Additionally, colleges and universities benefit from AP courses as they can identify students who meet their academic requirements.

The development of the AP Program began in 1951, with a pilot implementation in seven schools in 1952. The first common AP exam was administered in 1954. The program gradually expanded and became a national initiative in 1975, with 3,937 schools offering AP courses in 1976 and 6,720 schools in 1985. Over time, the number of students participating in the program also increased. In the 1990s, the federal government emphasized the importance of the program with George Bush's Education Goals 2000, introducing new technologies such as Videoconferencing to facilitate teacher training for AP courses.

AP courses cover six different subject areas: English, other languages, mathematics and computer science, science, social science and history, and art and music. As of 2000, there were 31 different AP courses available, which increased to 33 courses by 2005, the last year of analysis. The courses adhere to standardized curricula across the United States, and each course concludes with a standardized exam, serving as an assessment of students' abilities and academic performance. Annually, a committee comprising professors and secondary school teachers reviews and updates the existing curriculum and exams if necessary. The committee also sets grading standards to ensure that AP scores accurately reflect college-level performance (Agency. 2006)). AP exam scores are converted into grades ranging from 1 to 5. The examination fee was $72 in 1995 and increased to $83 by 2005.

Towards the end of the fourth decade of the AP Program, local school systems, education agencies, and foundations started providing incentives to encourage greater participation in the program. These incentives targeted both teachers and students. Teachers received reimbursements on an hourly or semesterly basis for teaching AP courses, while students received reimbursements for enrolling in AP courses, passing AP exams, or achieving top scores. In 1995, Texas initiated the Advanced Placement Incentive Program, and several other states, including Florida, Georgia, and North Carolina, followed suit (Rothschild 1999).

2.2 Advanced Placement Incentive Program

The Texas AP Incentive Program (APIP) was established in 1993 by the 73rd Texas Legislature with the purpose of acknowledging and rewarding students, teachers, and schools that achieve the educational objectives set forth by the state (Code 1996). The program offered various awards and subsidies, including teacher-targeted incentives like reimbursement of up to $450 for attending subsidized teacher training, and student-targeted incentives such as exam fee waivers. Additionally, school-targeted incentives were introduced in 2000, encompassing equipment grants and $100 per student who achieved a score of 3 or higher in the AP examination.

APIP funding distribution commenced in 1997, with the total funding for the State of Texas amounting to approximately $1,000,000 until the year 2000. Subsequently, in 2000, the contribution from the Foundation School Program (General Appropriations Act, Article III, Rider 30, and Strategy B.1.1, 76th Legislature) escalated significantly from $500,000 to $19,000,000 due to the

Table 1: Texas AP-IB Incentives

Incentive Target	Incentive Description	Funded Since 1994-95 Biennium	Funded in 2000-01 Biennium
School	A one-time $3,000 equipment grant for providing a college-level Advanced Placement (AP) & or International Baccalaureate (IB) course to be paid to a school based on need as determined by the commissioner.	No	Yes
School	$100 for each student who scores a three or better on a college-level AP examination or four or better on an IB examination.	No	Yes
Teacher	Subsidized teacher training, not to exceed $450 for each teacher, for a college-level AP or IB course.	Yes	Yes
Teacher	A one-time award of $250 for teaching a college-level AP or IB course for the first time.	No	No
Teacher	A share of the teacher bonus pool, which shall be distributed by the teacher's school in shares proportional to the number of courses taught. Fifty dollars may be deposited in the teacher bonus pool for each student enrolled in the school who scores a three or better on an AP examination or four or better on an IB examination.	No	No
Student	A student receiving a score of three or better on an AP examination or four or better on an IB examination may receive reimbursement, not to exceed $65, for the testing fee.	No	No
Student	The agency may pay for all AP and IB examinations taken by students who take a PEIMS-designated AP/IB course in the subject of the test.	No	Yes
Student	Students in financial need will receive further federal and state fee reductions.	Yes	Yes

Data Sources: TEC §§28.052-28.054 and Rider 30 of the Appropriations Act, Article III - Education, 76th Texas Legislature.

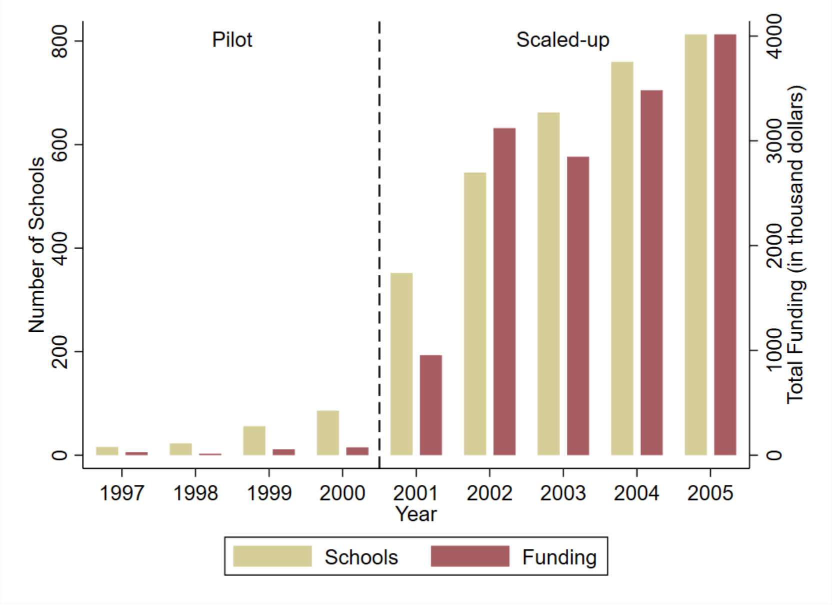
TEA correspondence from the commissioner dated 8/26/99 can be seen at http://www.tea.state.tx.us/taa/aas990826.html; dated 12/10/99 at http://www.tea.state.tx.us/taa/cur991210.html; and dated 3/22/00 at http://www.tea.state.tx.us/taa/gted000322.html.

implementation of No Child Left Behind. This substantial increase in funding facilitated the introduction of school-level incentives and higher exam fee waivers of up to 100%. For more detailed information regarding the incentives and changes in 2000, refer to Table 1.

During the initial years of the program, APIP funding was primarily sourced from AP Strategies, a nonprofit organization based in Dallas, TX. AP Strategies collected funds from private donors, with private donors covering 60% to 75% of the program's costs, while the remaining expenses were borne by the respective districts. However, with the expansion of the program following the introduction of No Child Left Behind funds in 2000, districts were able to cover the entire cost of APIP in each school, resulting in a decreased reliance on AP Strategies for funding. Eventually, AP Strategies were disbanded in 2012.

Prior to 2000, schools seeking APIP funding had to apply through AP Strategies and be placed on a waitlist. Private donors were then given the liberty to choose which school or schools they wished to fund from the waitlist. Upon a donor's agreement to fund a school, the curriculum creation and teacher training process would commence, and the program would be implemented in the follow-

Figure 1: The Program Implementation and Total Funding Over Time



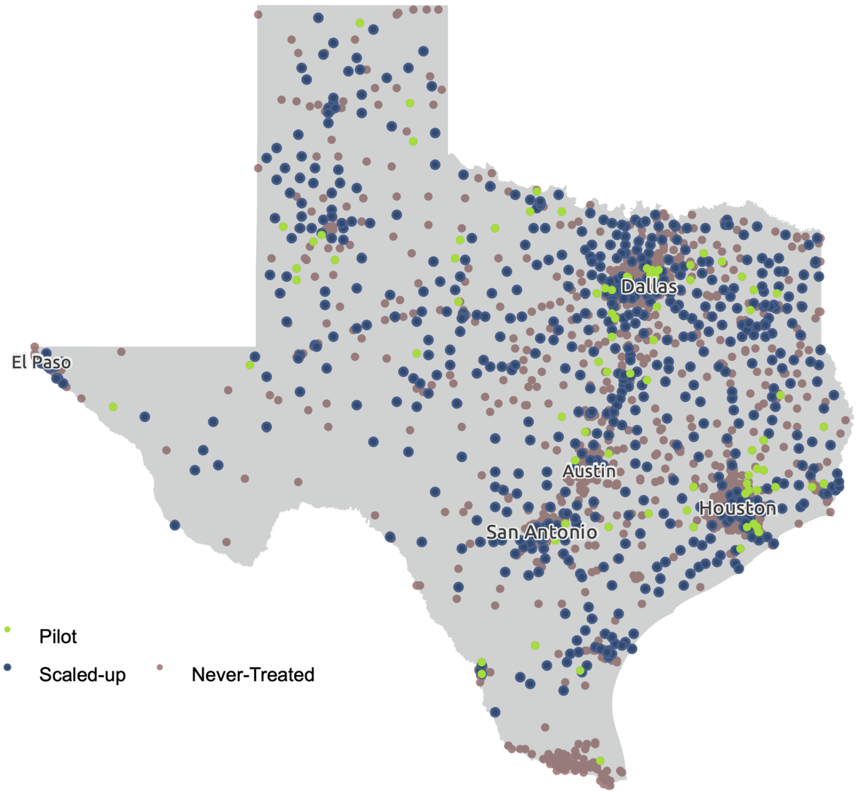
Notes: The figure presents raw total number of schools which have funding through APIP program in green, and raw total funding of APIP provides in blue for each secondary school year I use in the analysis. 1997 means 1996-1997 academic year. The dashed vertical line shows the time when the program is scaled up.

ing calendar year. Since donors could select the school or district to fund, and the number of donors was less than the number of applicants, donor preferences and availability created a timing variation in the program's implementation (Jackson 2010)

The following figure illustrates the evolution of APIP over time. The left-side y-axis represents the total number of secondary schools receiving APIP funding, showing a significant increase from 31 schools in 2000 to 400 schools in 2001. Similarly, the total funding allocated to the program experienced a parallel surge from $75,000 in 2000 to $1,000,000 in 2001, as presented on the right-side y-axis. The dashed vertical line shows the time when the program is scaled up and the increase after that time is clear.

The following map presents the locations of high schools separated by their treatment status over the state of Texas. the visual evidence shows that the distribution of the schools do not follow a specific rule, such as pilot or scaled-up schools are located close to each other, or a treatment arm is located in city center or college towns.

Figure 2: Location of High Schools in Texas



3 Data

The present study makes use of a longitudinal dataset that has been meticulously collected and curated by the Texas Education Research Center (ERC). This dataset comprises diverse student-level and school-level information, affording the ability to track comprehensive data concerning public schools and their students, spanning from kindergarten to 12th grade, acquired from the Texas Education Agency (TEA). Additionally, it includes data on post-secondary private or public schools and their students, provided they were previously enrolled in any K-12 school, sourced from the Texas Higher Education Coordination Board (THECB). Furthermore, the dataset encompasses longer-term post-education outcomes, obtained from the Texas Workforce Commission (TWC). As a result, the study benefits from information regarding the entire population of Texas secondary school students and their respective long-run outcomes.

Specifically, the TEA data inventory encompasses pertinent student-level variables, including the courses each student undertakes, the semesters in which these courses are taken, whether they pertain to AP courses, course pass rates, attendance records, high school graduation status, and demographic information. Additionally, the dataset contains school-level variables, such as annual school budgets, funding received from various programs and its respective amounts, teacher composition and demographics, among others. This information is of utmost significance for the analysis, as it enables the identification of schools that received funding and the corresponding years, which serves as the basis for identification variation. Furthermore, it facilitates the observation of first-stage outcomes, such as AP enrollment and the number of AP courses offered in each school year. THECB data includes students' college and university applications, outcomes of these applications, and the courses pursued throughout their college education. TWC data provides insight into quarterly wages and employment sectors.

The study utilizes variables that are available for students and schools from 1993 to the present day. The first instance of APIP funding received by a secondary school occurred in 1997. A 4-year period preceding the program's implementation allows for controlling pre-trends. The dataset's terminal year for secondary school analyses is 2005, allowing for the examination of long-run outcomes related to college, wages, and employment. The main sample encompasses all students enrolled in a secondary school between 1993 and 2005, with their long-run outcomes spanning until 2020. This dataset comprises 122,233,258 observations, with 5,211,896 unique student IDs for secondary school analyses. For college-level outcomes, the number decreases to 2,779,212 unique student IDs, with approximately 100,000 unique student IDs dropped due to reasons such as missing or inconsistent data.

Various outcome measures have been devised, including total AP enrollment, total number of AP courses offered by each school, number of college enrollees, number of college graduates per secondary school, total loans and scholarships obtained by students in college (conditional and unconditional on college enrollment), and total wages earned by individuals 1 to 15 years after secondary

Table 2: Treatment and Control Groups Summary

School Groups	Implementation Time	Number of Schools
Pilot	1997-2000	87
Scaled-up	2001-2005	731
Counterfactual	Either after 2005 or never implement	1982

school graduation, with yearly wages adjusted to 2020 dollars.

As for control variables, school-level time-invariant variables have been created, incorporating information on race, economic conditions, and demographic characteristics. These variables encompass the percentage of students considered at risk, gifted, in special education, belonging to specific racial or ethnic groups (Black, Hispanic, Asian, or Native American), enrolled in free or reduced meal programs, or experiencing other economic disadvantages. Each variable has been formulated by calculating the number of students in each category for each school and year, divided by the total number of students within that school and year.

4 Empirical Framework

The identification strategy employed in this study is based on exploiting the temporal variance in the implementation of the program. Due to limited donor resources, the APIP could not be simultaneously implemented in all interested schools until 2001. This circumstance allows for the restriction of the estimation sample to schools that either implemented or were scheduled to implement APIP by 2001. Schools that did not implement APIP until 2005 serve as the counterfactual group to assess the change in outcomes for the pilot schools. For the scaled-up schools, the difference in implementation timing arises from variations in their application to the program. Consequently, the scaled-up group comprises schools that implemented APIP between 2001 and 2005, while the counterfactual group consists of schools that did not implement APIP until 2005. Each treatment arm exhibits staggered adoption, and the counterfactual groups remain the same for both pilot and scaled-up schools. The estimation method employed is a staggered difference-in-differences (DiD) framework, where outcomes are estimated between schools that implemented APIP and those that did not. Table 2 summarizes which schools are included in which group of treatment or control groups.

Standard estimation approach for such a scenario commonly employes the Two-way Fixed Effects (TWFE) estimator. However, recent literature has highlighted the limitations of TWFE when dealing with groups treated at various time points (Goodman-Bacon 2021; Callaway and Sant'Anna 2021; Sun and Abraham 2021). TWFE uses already-treated groups as controls, overlooking potential dynamic treatment effects. The APIP setting is possibly have a dynamic effect, because the longer a schools have the funding the longer the school offers more AP courses. This ends up with a more experienced, better teaching skills for the teachers and the effect size can get bigger over time. Even though there is no dynamic effect, TWFE is still introducing bias when there are heterogeneous treatment effects across groups. It is reasonable to expect heterogeneous treatment effect across groups because of the reasons such as earlier adoption might signal a higher motivation. To address these issues, this study adopts the DiD method proposed by Callaway and Sant'Anna (Callaway and Sant'Anna 2021), which accommodates multiple time periods, variation in treatment timing, and parallel trends after conditioning on observable characteristics.

Four assumptions of Callaway and Sant'Anna DiD method are presented formally: First, the data used for estimation is panel or repeated cross-sectional data. Second, the treatment should only be turned on. Third, parallel trends should be maintained conditional on covariates. Fourth, the treatment and counterfactual groups should have units with approximately the same propensity score. The first assumption is satisfied by the longitudinal nature of the unique dataset ERC. With ERC, I can follow outcomes and treatment status of students, schools and districts over time. The second assumption is met due to the nature of APIP, as schools do not need to apply for the program annually. Once they applied and approved, the schools remain treated, i.e., continue to get the APIP funding every year.

There are potential challenges to identification, as school characteristics might influence the timing of APIP application. Callaway and Sant'Anna's method differs from other recent difference-in-differences literature (AAthey and Imbens 2022; Borusyak, Jaravel, and Spiess 2022; De Chaisemartin and d'Haultfoeuille 2020; Sun and Abraham 2021) in that it attempts to minimize parallel trends assumptions and allows for flexible incorporation of covariates. This is achieved through conditional parallel trends, the third assumption, where the parallel trends assumption holds after conditioning on some covariates X. To achieve this, the method employs a matching propensity score approach to assign more weight to similar schools, thereby relaxing the canonical parallel trends assumption. Throughout the paper, both the Callaway and Sant'Anna method with and without matching propensity score are utilized to address potential issues related to timing differences in APIP application. The fourth assumption is assured by not conditioning on a large set of covariates X. The probability to be in a certain group is going to get higher with the set of X gets larger.

Following Callaway and Sant'Anna, in the first stage of analysis, I start with estimating the propensity score to ensure that the conditional parallel trend assumption holds. By doing that, I choose schools from never-treated schools that are observationally similar to each treatment arm pilot and scaled-up schools, thus relax canonical parallel trend assumption. Since there are multiple treatment dates for multiple groups, a unique propensity score is calculated for each group. For each $ATT(g, t)$, a generalized propensity score is calculated in the following way:

$$\widehat{p(x)} = P(G_g = 1|X, G_C + C = 1) \tag{1}$$

where g is the is the point in time where the group is first treated (i.e. 1997 for the schools who get funded in 1997), G_g is a binary variable which is equal to one if a school is first treated in year g, C is a binary variable which is equal to one if a school is never-treated. $p(x)$ is the probability of being in a certain group conditional on pre-treatment covariates X. Pre-treatment covariates of percentage of students being at risk, needing special education, being Black, being Hispanic, being enrolled into free meal or reduced meal programs before 1997, (i.e., before APIP is introduced) are used in the propensity score calculation instead of time-varying covariates.

Since the treatment is school-level, the pre-treatment observables used in matching process are constructed as school-level percentages, not student-level binary variables. To be consistent with the matching strategy, the outcomes are also constructed at school-level. Beside matching variables of percentage of demographics and economic indicators, outcome variables are constructed as follows: the number of AP courses, total AP enrollment, total number of students who enrolled in college, total number of students who graduated from college, employment rate, and mean wages in each school.

Table 3 Panel A summarizes the statistics for pilot high schools and scaled-up high schools in the sample. Column 1 shows the mean characteristics for pilot high schools, column 2 shows the means of control group used as counterfactual for pilot schools, column 3 the mean characteristics for scaled-up schools, column 4 shows the means of control group used as counterfactual for scaled-up schools column 5 shows the means of control group before the matching propensity score. Table 3 shows that pilot and scaled-up schools have similar percentage of students at risk, special education or black students. While, scaled-up schools have more students enrolled in free meal and there are also more Hispanic students. These results reported by using a matching propensity score to mimic the first step of Callaway & San'Anna DiD method (Callaway and Sant'Anna 2021). Appendix figures **??** shows the density functions of propensity scores for each treatment arm and their comparison groups.

Table 3 Panel B shows the pre-treatment means of covariates that are not used to match on for pilot schools, scaled-up schools, their correspondent control groups and the all sample of never-treated schools. The pre-treatment time-invariant covariates that are not used in the first stage of the Callaway & Sant'anna estimation, i.e., estimating propensity scores to determine control groups with respect to treatment groups are similar between each treatment arm and its respective control group. Means of all never-treated schools are included to convince the reader the matching variables improve the similarity of non-matched outcomes.

Then, I estimate group-time average treatment effects as follows:

Table 3: Summary Statistics

Panel A: Matching Variables

	Pilot	Matched Control	Scaled-up	Matched Control	Control
At Risk (%)	.2959	.3058	.3183	.3206	.4367
		[0.626]		[0.709]	
Special Education (%)	.1005	.1181	.0987	.0929	.1780
		[0.570]		[0.430]	
Black (%)	.0721	.0636	.0738	.0728	.1224
		[0.392]		[0.800]	
Hispanic (%)	.1265	.1209	.1983	.2077	.2543
		[0.745]		[0.229]	
Free Meal (%)	.1735	.1901	.2020	.2087	.2831
		[0.385]		[0.269]	

Panel B: Non-matching Variables

	Pilot	Matched Control	Scaled-up	Matched Control	Control
Median Income	34579	37047	34807	37679**	36868
		[.105]		[.021]	
% City	.120	.167	.211	.190	.173
		[.159]		[.291]	
% Not Econ Disadv	.629	.560	.560	.538	.478
		[.099]		[.194]	
Total Number of st	1192	986*	1250	1244	608
		[.077]		[.102]	
AP Course	4.500	4.170	4.016	3.775	3.582
		[.281]		[.189]	
AP Enrollment	109.674	98.013	80.747	78.852	46.256
		[.173]		[.752]	
College Enrollment	57.878	50.279*	39.281	39.244	24.578
		[.096]		[.985]	
College Graduation	52.006	45.667	34.640	34.201	21.746
		[.102]		[.837]	

Notes: Panel B present mean percentage of characteristics in secondary schools separated by two different treatment arms; pilot and scaled-up schools, control group, and control groups for each treatment arm after matched by propensity matching score. Panel A present differences between two different treatment arms; pilot and scaled-up schools and their control groups in pre-treatment unobservables to the matching process. Values in brackets show p-values for each matching result.

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\dfrac{\widehat{p(x)}C}{1-\widehat{p(x)}}}{E\left[\dfrac{\widehat{p(x)}C}{1-\widehat{p(x)}}\right]}\right)\left((Y_t - Y_{g-1}) - \hat{\mu}_{0,\Delta}(X)\right)\right] \tag{2}$$

where $\hat{\mu}_{0,\Delta}(X) = E[Y_t - Y_{g-1}|X, C = 1]$, g is the is the point in time where the group is first treated (i.e. 1997 for the schools who get funded in 1997), t is the year, G_g is a binary variable which is equal to one if a school is first treated in year g, C is a binary variable which is equal to one if the school is never-treated, and $\widehat{p(x)}$ is the propensity score calculated by matching on the covariates listed in Table 3. $Y_t - Y_{g-1}$ shows the change between the year t and the year prior to the treatment, $g - 1$.

The equation shows the average treatment effect for treated group g in each time period t that generalizes the two-period parallel trends assumption to the case where there are multiple time periods and multiple treatment groups. The first part of the first term shows the weight calculated for each treatment group g, the second part of the forst term shows the weight for comparison group calculated by using propensity scores estimated in Equation 1, so the term presents the similarity between treatment and control group conditional on pre-treatment covariates X. The second term shows the difference in outcome before and after the treatment for the treated group g for each point in time t. The comparison time, i.e., before treatment, is the last period before the treatment starts for the group g.

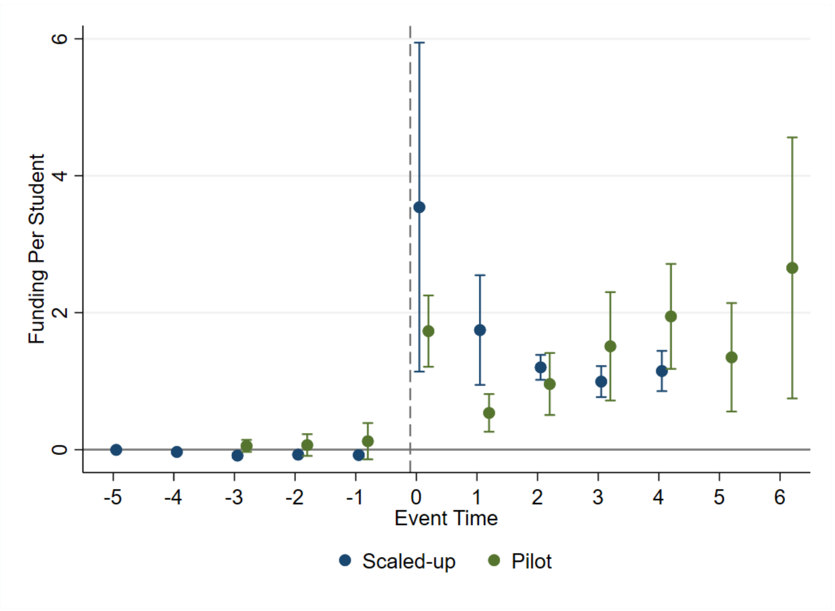
Following Callaway & Sant'Anna, I then aggregate the group-time average treatment effect to summarize:

$$\theta = \sum_{g \in G} \sum_{t=2}^{\tau} \omega(g,t) \times ATT(g,t) \tag{3}$$

where θ is the ATT which is calculated by getting simple weighted average of $ATT(g,t)$ estimated in Equation 2.

Since there are students who do not choose to enroll to an AP course even if the school implements APIP, one can consider the coefficient of interest θ corresponding to the Intention-to-treat (ITT) effect or if the outcomes are analyzed at the school-level, it can be considered as average treatment effect (ATT). The strategy relies on the assumption that schools with APIP would behave the same as schools which applied APIP but did not find a donor to fund the program for them or did not apply yet. Although this assumption is not directly testable, Table 3 shows that observable characteristics of pilot and scaled-up schools are similar. The estimations are made separately for pilot and scaled-up schools. Additionally, the study visually presents trends in outcome variables in an event study for both pilot and scaled-up schools to explore parallel trends and dynamic effects. The specifications of the equation for visual evidence of parallel trends and dynamic effects are as follows:

Figure 3: Event Study Estimates of the APIP Funding



Notes: This figure presents the ATT on APIP fundings with respect to the implementation time of APIP. Funding per student is calculated by total APIP funding per school over total number of students in that school.

$$\delta_{eP} = \sum_{g=1997}^{2000} \sum_{t=1995}^{2005} 1\{t-g=e\} P(G=g|t-g=e) ATT(g,t) \tag{4}$$

$$\delta_{eS} = \sum_{g=2001}^{2005} \sum_{t=1995}^{2005} 1\{t-g=e\} P(G=g|t-g=e) ATT(g,t) \tag{5}$$

where δ_{eP} and δ_{eS} are the average effect of the implementation of the program e periods after the treatment was implemented across all pilot schools in Equation 4 and across all scaled-up schools in Equation 5. The indicator function $1\{t-g=e\}$ restricts the effect to the identified group-time (g,t) average treatment effects, while $P(G=g|t-g=e)$ is the group-specific weight.

5 Results

Figure 3 shows the event study estimates for the APIP funding per student separated by pilot and scaled-up schools. The evidence for the assumption that there are parallel trends, is that the parameters for -3 to -1 for pilot schools and the parameters for -5 to -1 for scaled-up schools are all indistinguishable from zero. Callaway & Sant'Anna (Callaway and Sant'Anna 2021) also allows me to

Table 4: The Effect of APIP on AP Utilization

	Pilot		Scaled-up	
	(1)	Mean	(3)	Mean
Funding Per Student	1.32***	0.0	1.59***	0.0
	(.31)		(.51)	
Number of AP Courses	1.81***	4.50	.12	4.01
	(.47)		.14	
AP Enrollment	84.17**	163.84	8.21	100.81
	(40.55)		8.20	
AP Students	57.84**	109.42	4.78	80.74
	(25.35)		(5.55)	
Number of AP Passes	76.77**	143.96	5.368	85.76
	(32.70)		(8.092)	
School Controls	Yes		Yes	

Notes: This table presents the change in the funding per student, the number of AP courses and AP enrollment after implementation of APIP. Columns show total change and change in different treatment arms, pilot and scaled-up schools along with their pre-treatment means. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

test the conditional parallel trends assumption by a hypothesis test under the null hypothesis that is all pre-implementation coefficients are zero. In all cases, the test does not reject the null hypothesis with p-values of 0.45 and 0.39.

Table 4 shows the main results from the Equation 3 on the AP enrollment and the number of AP courses for the state of Texas and separately for pilot and scaled-up schools. Columns vary in the treatment arms and pre-treatment means of outcomes specified on the left. Results in which school characteristics such as the rate of students who are Hispanic or enrolled in free meal are controlled is posted, and the outcomes without school controls are not significantly different from the ones in Table 4.

Columns (1) and (3) shows the average effect of the implementation of APIP for pilot and scaled-up schools respectively to take the change in the program in 2001 into account. The outcome of funding per student is calculated by determining the total APIP funding per school and dividing it to the total number of students in that school. As shown in the first row, the increase in APIP funding does not differ much between treatment arms, the increase is \$1.32 for pilot schools and \$1.59 for scaled-up schools. The second outcome is the number of AP courses per school per year. This outcome counts each section as one, i.e., if there are two sections of Calculus I, the total includes both of them. The third outcome is the total AP enrollment per school, constructed by summing up all individual enrollments per school per year. If a student is enrolled in two different AP courses, then it is counted as two. In pilot schools,

Table 5: The Effect of APIP on Long-Run Outcomes

	Pilot		Main	
	(1)	Mean	(3)	Mean
Number of High School Grad	21.07	173.89	.047	146.71
	(12.97)		(3.539)	
College Enrollment	14.14***	57.87	-.21	39.28
	(4.86)		(1.70)	
College Graduation	2.94*	52.00	-.10	34.64
	(1.61)		(1.12)	
STEM Enrollment	.74**	6.67	.01	4.38
	(.31)		(.16)	
Non-STEM Enrollment	5.41***	48.07	.09	34.71
	(2.17)		(1.24)	
School Controls	Yes		Yes	

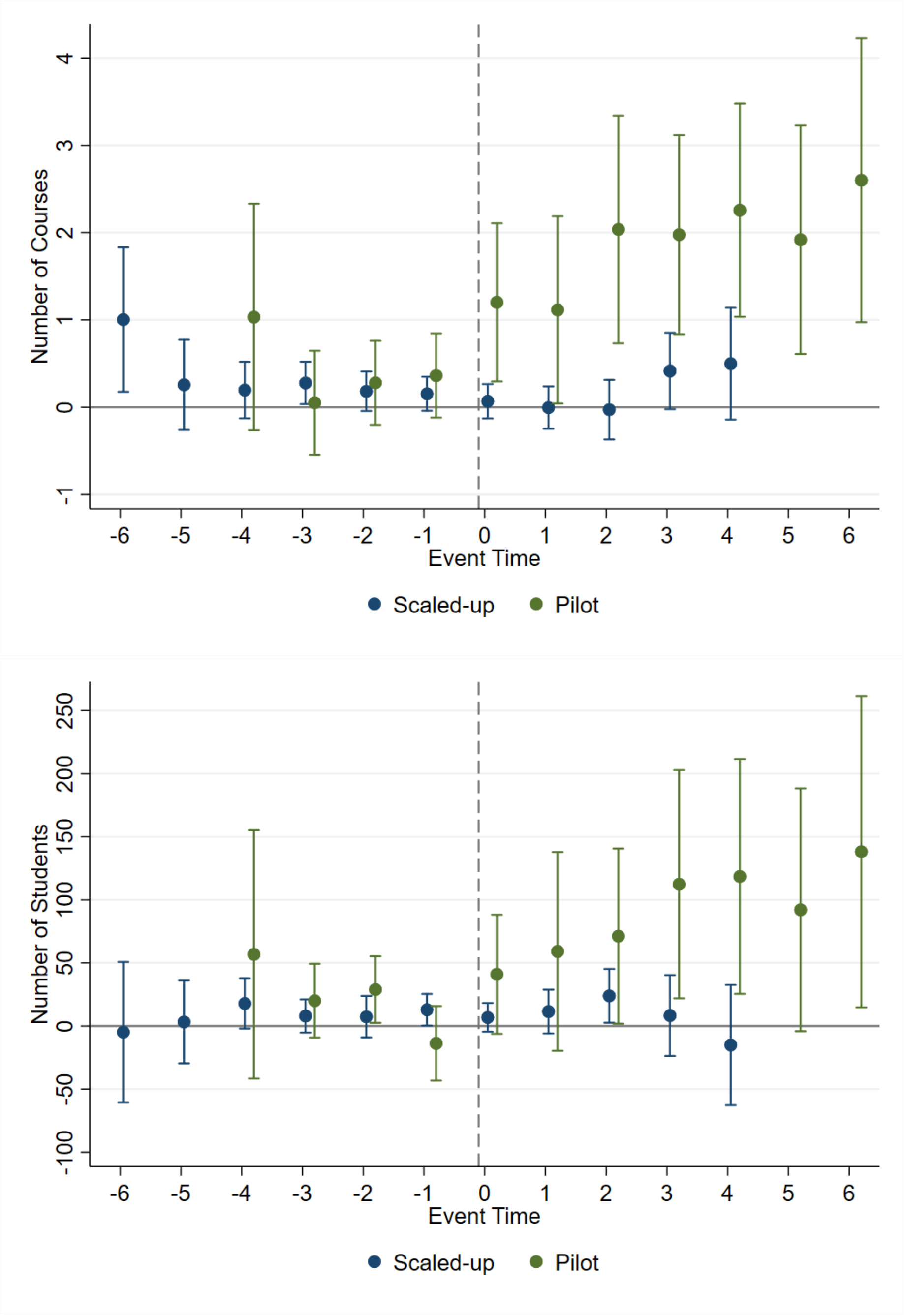
Notes: This table presents the change in the number of college enrolled students, the number of college graduated students and student loans per student unconditional of college enrollment status after implementation of APIP. Columns show the change in different treatment arms, pilot and scaled-up schools along with their pre-treatment means. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

the increase in the AP enrollment is around %50 (84 students) and the increase in the number of AP courses %40 (1.8 courses). The effect is much more smaller and is not statistically different from zero for scaled-up schools. AP students outcome is constructed as the new enrollment of a unique student. It is the sum of all students who are enrolled in at least one AP course. In pilot schools, the increase in AP students is %53 (57 students).

The positive dynamic effect in the number of AP courses and AP enrollment after the implementation of the program can be seen in Figure 4. For almost all years after the implementation, the effect gets larger over time. Callaway & Sant'Anna (Callaway and Sant'Anna 2021) parallel trends assumption test does not reject the null hypothesis with p-values of .31, .10 (Fund per student), .13, .11 (AP course), .15, .27 (AP enrollment), .41, and .38 (AP students). The results for TWFE and Sun & Abraham Sun and Abraham 2021 can be found Appendix Figure 11. While TWFE estimator overestimates the results, Sun & Abraham DiD estimator works worse for the parallel trends assumption. The robustness figures are reassuring that our estimator works the best in our setting.

Long run effects of the program such as on college enrollment and on wages are important for several reasons. First, they help to understand if the financial burden of APIP pays off to the government. Second, it is valuable to learn for a high school student that what positive effects a student can get from enrolling an AP course. Table 5 shows the results for high school graduation, college

Figure 4: Event Study Estimates of the Number of AP Courses and AP Enrollment



Notes: These figures present event study estimates of the effect of APIP on the number of AP courses and AP enrollment separated by pilot and scaled-up schools. School controls are included. Standard errors are clustered at district level.

enrollment, college graduation, and major choices summarized as STEM or Non-STEM enrollment. Columns (1) and (2) display outcomes for pilot schools with school characteristics controls and their means before the treatment. Columns (3) and (4) display outcomes for scaled-up schools with school characteristics controls and their means before treatment.

There is no statistically significant change in high school graduation. This is caused by following: the top students are affected by AP incentives. Thus, it is very likely that APIP does not change any student's high school graduation prospect. College enrollment and college graduation increase by %25 and %6 for pilot schools, while they do not change for scaled-up schools. There is not a significant evidence of change in major choices as seen from the increases in STEM and non-STEM enrollment, enrollment increases by %11 in both STEM and non-STEM majors.

The event study results estimated by the Equations 4 and 5 and using number of college enrollees and the number of college graduates an outcomes are presented in the Figure 5. These two figures are in line with the previous event studies where the dynamic effect is easily seen for pilot schools while the effect on scaled-up schools are null in all years after the implementation of APIP.

The chosen labor market outcomes are employment rate between ages of 25-30, and the natural log of mean wage between ages of 25-30. The employment rate is calculated as the number of students being employed between ages of 25-30 over the total number of students in a school. The log mean wage is calculated by adjusting all yearly wages to 2020 U.S dollars, calculating the mean of the wages between 25-30 for each student and then taking the average over school and graduation year. The sample is restricted to the high school graduates since it is expected that better students are affected from an AP increase more. The same outcomes with the whole sample, i.e. not conditional on high school graduation, are lower than the outcomes for the sample conditional on high school graduation and can be found in the Appendix.

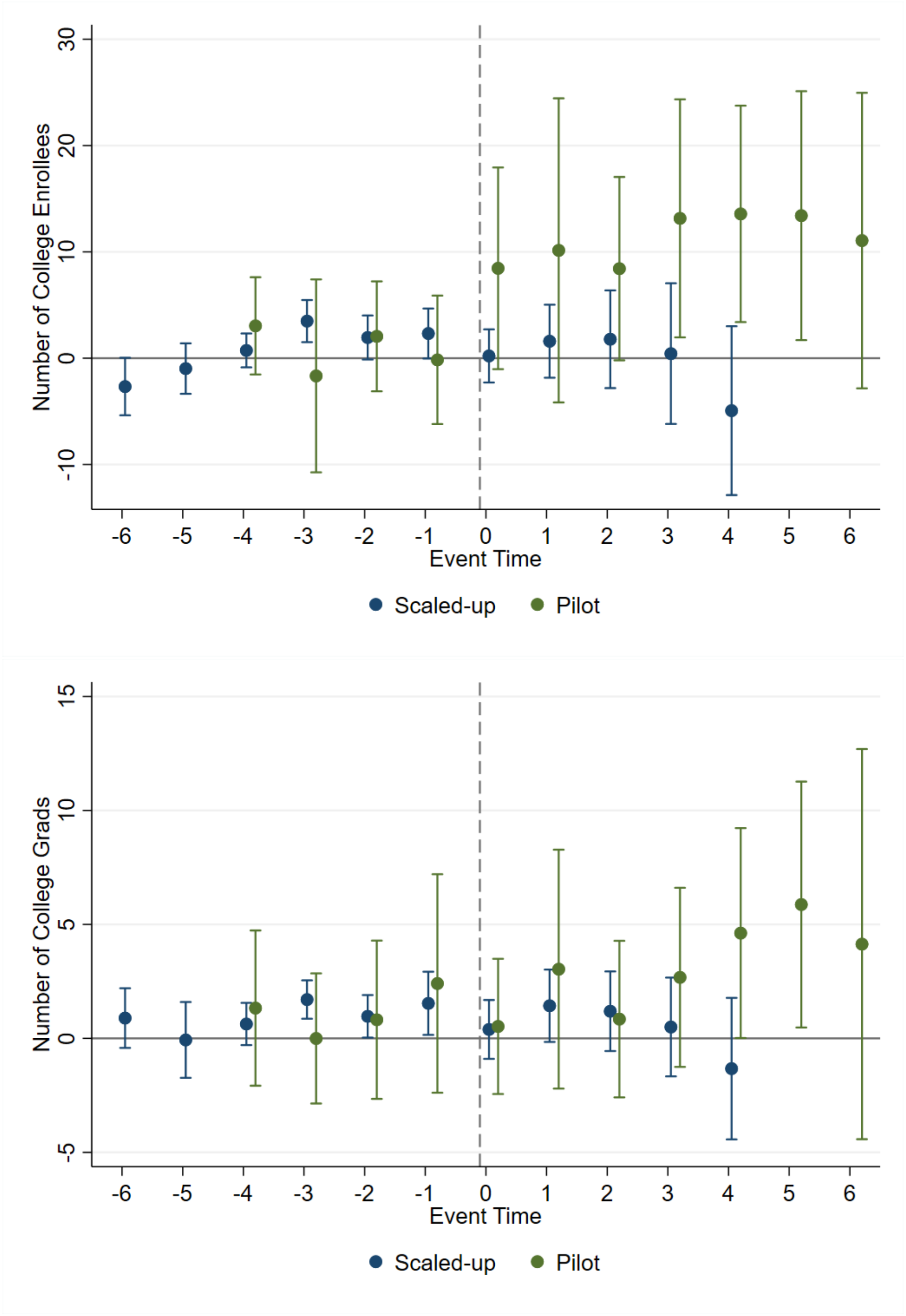
Table 6 shows the ATT of APIP on employment rate and log wages between ages of 25-30 conditional on high school graduation. Employment rate increases by 4 percentage point while wages increase around %16 percent for the ages of 25-30.

The event study results estimated by the Equations 4 and 5 and using employment rate and the natural log of mean wages between the ages of 25-30 as outcomes are presented in the Figure 6. These two figures are in line with the previous event studies where the dynamic effect is easily seen for pilot schools while the effect on scaled-up schools are null in all years after the implementation of APIP.

5.0.1 Courses Introduced

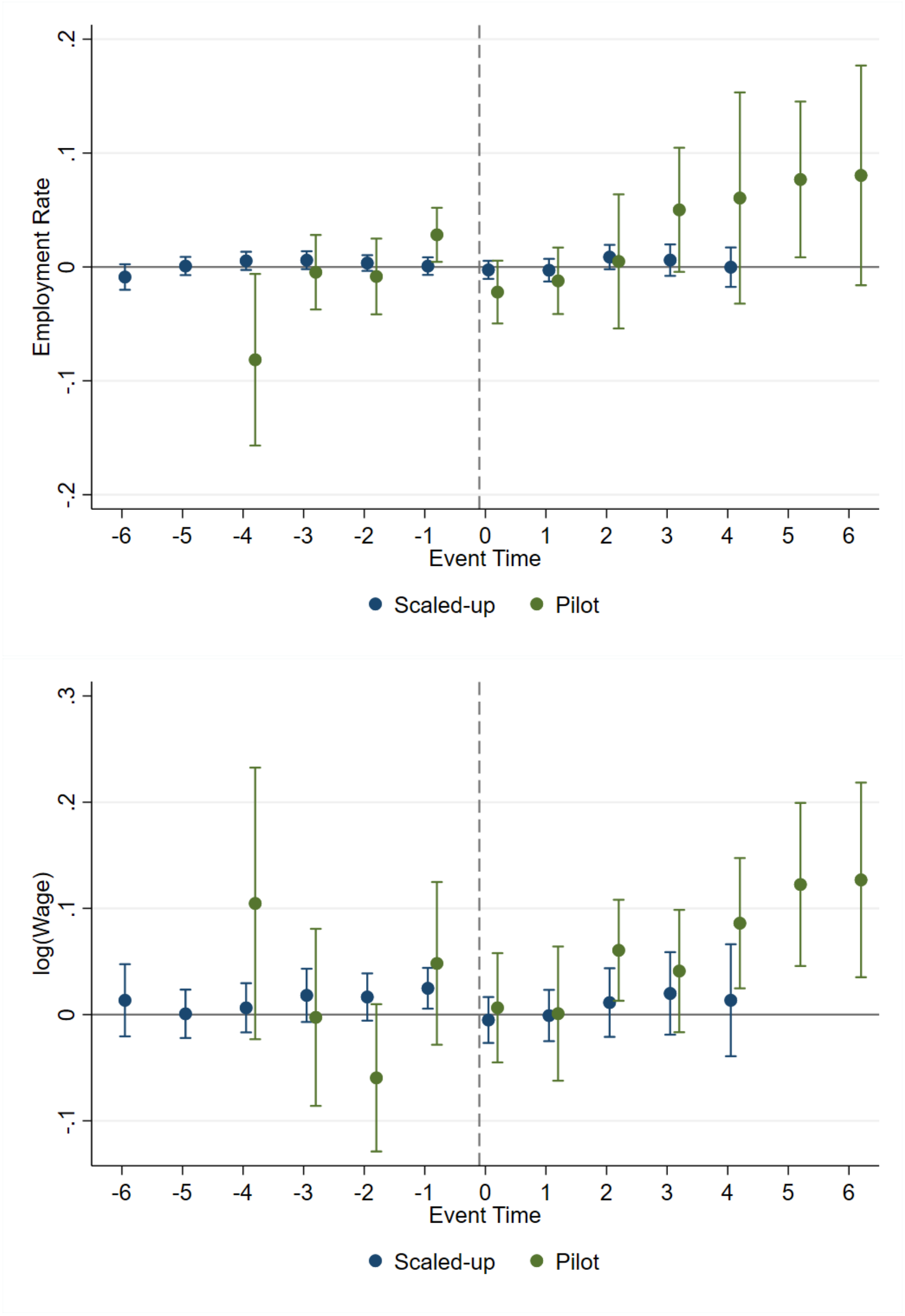
The introduction of AP courses was approximately 30 years earlier than the introduction of APIP. Since the AP courses are available for a long time, many schools have already been offering those courses. The results in Table 7 show that the existence of APIP help secondary schools to broaden their AP course

Figure 5: Event Study Estimates of College Outcomes



Notes: These figures present event study estimates of the effect of APIP on the number of students who enroll in a 4-year college and on the number of students who graduated from a 4-year college separated by pilot and scaled-up schools. School controls are included. Standard errors are clustered at district level.

Figure 6: Event Study Estimates of Labor Market Outcomes



Notes: These figures present event study estimates of the effect of APIP on the employment rate and on the natural log of mean wages between the ages of 25-30 separated by pilot and scaled-up schools. School controls are included. Standard errors are clustered at district level.

Table 6: The Effect of APIP on Labor Market Outcomes

	Pilot		Scaled-up	
		Mean		Mean
Employment Rate	.015	.773	.002	.786
	(.010)		(.004)	
log(Wage)	.069***	$29,611	.005	$29,179
	(.024)		(.012)	
School Controls	Yes		Yes	

Notes: This table presents the change in the employment rate and the log wage between the ages of 25-30 after implementation of APIP. Columns show the change in different treatment arms, early and late-treated schools along with their pre-treatment means. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

selection. Many secondary schools offered AP Math and AP English, so APIP does not affect the number of those courses, while it increases AP Science, AP social Science, AP Foreign Languages and AP Art to provide students a variety in the AP course list to increase AP utilization.

6 Potential Mechanisms

The implementation of APIP increases the number of AP classes by 2 and the AP enrollment by 98 in pilot schools, while the effect is small and insignificant for scaled-up schools. My focus is to understand how pilot schools differ from scaled-up schools to be able to utilize the APIP better. I focus on four potential mechanisms: locational differences, motivational differences, hawthorne effect, and supply and demand side mechanisms signaled by the difference in AP teachers and the difference in AP classroom sizes.

6.1 Location

It is possible to think if the pilot schools are geographically close to each other or they are close to college towns and students in those schools are more aware of the opportunities they will have if they take AP courses. In other words, the location hypothesis says that pilot schools are located in specific areas such as college towns or city centers. Figure 2 shows the pilot and scaled-up schools compared to never-treated schools in the dataset. And, it can be easily seen that the distribution of the schools over the state of Texas is pretty uniform. Figure 17 presents the densities of urbanicity level available in ERC, which are rural, town, suburb, and city for pilot, scaled-up and never-treated schools. The distributions are similar. Moreover, the outcomes are robust to the analyses includes urbanicity level as a matching variable.

Lastly, Table 8 presents the heterogeneity of the main results by urbanicity levels. For pilot schools, AP courses increases in all levels of urbanicity, and AP enrollment increases in suburbs and cities. The main reason AP enrollment

Table 7: Number of Courses Separated by Subject Areas

	Pilot	Scaled-up
English	.061	-.016
	(.098)	(.029)
Math	.043	.001
	(.041)	(.012)
Science	.687***	-.039
	(.120)	(.037)
Social Science	.300***	.101**
	(.099)	(.040)
Foreign Lang	.302**	-.035
	(.150)	(.053)
Art	.479***	.122**
	(.132)	(.056)

Notes: The table presents the effect of APIP on the number of AP courses in different subject areas, separated by treatment arms. The results are in nominal terms. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

is not significant in rural regions and towns is the low power in those regions because of the cohort sizes as shown in the table. If the location is actually the reason behind the difference between pilot and scaled-up schools, we need to expect similar increase in AP enrollment in suburbs and cities for scaled-up schools too. However, for scaled-up schools, there is an increase in AP courses only in the cities, while there is no effect on AP enrollment. With all of these evidence, it is clear that the location of the schools is not the main driver of failure of the scaling-up.

6.2 Motivation

One can think that since schools need to apply for the APIP funding, highly motivated schools are early-adopters of the program. In other words, pilot schools applied for APIP funding earlier than scaled-up schools because pilot schools are more motivated to increase AP utilization. To test this hypothesis, I utilize one of the unique features of Callaway & Sant'anna estimation which is to aggregate group-time ATTs over time to have group-specific ATTs. If the motivation hypothesis is valid, we need to expect higher AP utilization for the groups which get the funding earlier.

The Figure 7 presents the ATT of Number of AP courses and AP enrollment by groups. The left y axis shows the number of courses, while the right y axis shows the number of AP enrollment. The two outcomes are positively correlated with each other. Moreover, the figure shows that I can eliminate high motivation hypothesis since the outcomes are not getting smaller over the groups. Instead, it is similarly high in pilot groups, while they are similarly low in scaled-up

Table 8: Heterogeneity by Urbanicity Levels

Panel A: Pilot Schools

	Rural	Town	Suburb	City
AP Courses	1.17*	1.27**	1.77**	2.90***
	(.60)	(.53)	(.94)	(1.45)
AP Enrollment	41.95	-8.54	101.72**	235.67***
	(31.29)	(21.60)	(45.91)	(60.16)

Panel B: Scaled-up schools

	Rural	Town	Suburb	City
AP Courses	.29	.13	-.23	1.28***
	(.25)	(.38)	(.69)	(.40)
AP Enrollment	20.17	6.17	22.47	36.07
	(12.28)	(11.43)	(36.02)	(24.36)

Notes: The figure presents DiD analysis for pilot and scaled-up schools, separated by their urbanicity level. The results are in nominal terms. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

Figure 7: AP Utilization Outcomes by Group

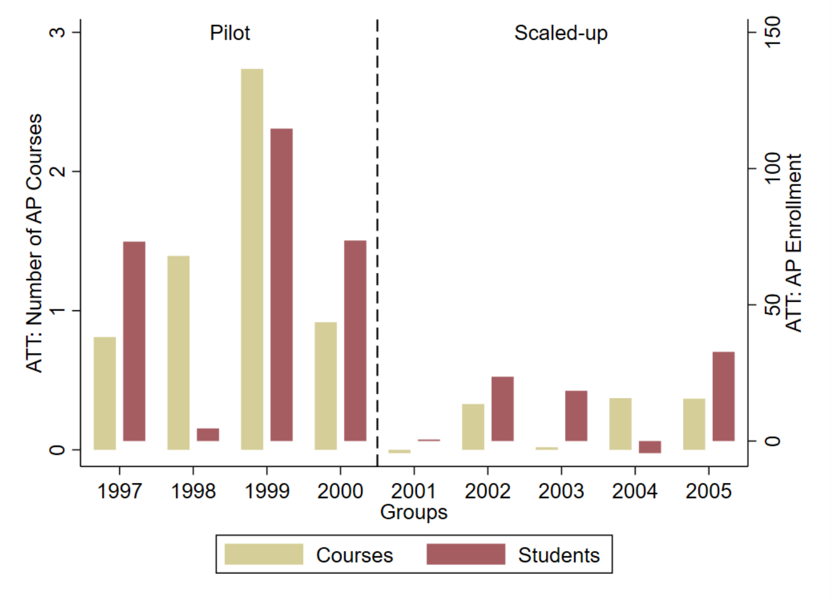


Table 9: Heterogeneity by the number of Pre-Period AP Courses

	Pilot		Scaled-up	
	<Median	>Median	<Median	>Median
Total AP Courses	-.08	1.32**	.07	.07
	(.24)	(.55)	(.09)	(.17)
AP Enrollment	32.91	70.51**	1.07	7.98
	(36.95)	(33.42)	(19.51)	(8.67)

Notes: The table presentes the effect of APIP on the number of AP courses and AP enrollment separated by treatment arms and their status with respect to the pre-treatment median number of AP courses in Texas. Median is defined as the median number of AP courses of all sample before 1997, and equal to 2. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

groups.

6.3 Hawthorne Effect

In psychology, the Hawthorne Effect is defined as changing the behavior in the existence of an observer. Since the number of schools which adopted APIP are low and the funds are collected by a local NGO called AP Strategies in pilot phase, it is easier for AP Strategies to track the changes in pilot schools. The organization might be interested in observing higher AP Utilization to see if the low budget is allocated efficiently. On the other hand, the schools are unlikely to be observed for the funding in the scaled-up phase since it is harder to keep track of each school and their spending federally. The main idea of this mechanism is since there is an observer, i.e. the NGO, pilot schools try hard to increase AP utilization. To test this hypothesis, I classify high schools according to their position with respect to the median number of AP courses in Texas. The reason behind the test is as follows: if there is an observer effect, then we need to expect similar increase in AP utilization for all subsamples of pilot schools. As easily seen in Table 9, there are positive and significant results for pilot schools which have more AP courses that the median number of AP courses before the implementation of APIP. The analysis presents that there is a heterogeneity among pilot schools by the number of AP courses offered before APIP introduced, the schools which offered more AP courses increase AP utilization more with APIP. On the other hand, there is no difference for scaled-up schools according to their AP course offerings before the introduction of APIP.

6.4 Demand for & Supply of AP Program

Table 9 shows that among the pilot schools, the schools which already offer more AP courses are more successful to construct new AP courses and attract students to take them. These results direct us to a possible demand or supply side mechanism. We cannot conclude that these results present the underlying mechanism is the demand for the AP program, since it does not explain the differences in outcomes for pilot and scaled-up schools. However, it can signal that we can search more evidence of demand and supply mechanisms, since the schools increase AP utilization more already have more AP courses. On one hand, the students may demand more AP courses so that these schools offer more AP courses. By this, we can find an underlying demand mechanism. On the other hand, schools may supply more AP courses to make their students more successful, more competitive and more college-ready. By this, we can find an underlying supply mechanism. To investigate these mechanisms more, I look at the compositional changes in AP teachers and AP classroom sizes.

6.4.1 AP Teachers

An approach to see if the increase in pilot schools is driven by a supply change in those schools is looking for changes in the number pf AP teachers and in the characteristics of teachers. If there is an increase in the number of teachers, it means there are more teachers who are willing to teach AP courses. If there is an improvement in the education levels of AP teachers, it means that either schools hire more educated teachers with the APIP fund or teachers spend the fund for their own education. The results could signal a supply-side mechanism leading to positive outcomes for pilot schools. However, Table 10 shows that none of these outcomes changed in any type of the schools with APIP. The results are obtained by using Equations 4 and 5 where the variables on the left side are used as outcome variables in the equations.

6.4.2 AP Classroom Sizes

Lastly, one could think that there might be a difference in the AP classroom sizes in those groups. If classroom size is higher in pilot schools, it signals that there is a higher demand for AP courses in pilot schools. To check if this is the case, I look at the number of students per AP classroom and compare pilot and scaled-up schools by also taking the number of AP courses offered into account. Figure 8 shows that the number of students per AP classroom are lower in scaled-up schools. This fact does not change with the implementation of the APIP, even though the total number of AP courses offered is lower in scaled-up schools while is increased in pilot schools.

Furthermore, if the AP classroom sizes are the reason behind the difference between pilot and scaled-up schools, then we need to expect a higher AP utilization for the scaled-up schools which have a higher AP classroom size. To test this, I split the sample according to pre-period median AP classroom sizes. The median AP classroom size is 32. Below median means the schools which have mean number of AP classroom size is less than 32 and above median is more than 32. Table **??** presents the results for splitted samples of pilot and scaled-up schools. Although APIP funding per student increases in all types of schools with any AP classroom size, the increase in AP enrollment occurs

Table 10: Changes in AP Teachers Characteristics

	Pilot	Scaled-up
Number of AP Teachers	.586	.287
	(.487)	(.218)
Asian	-.012	.002
	(.011)	(.010)
Hispanic	.082	.039
	(.056)	(.036)
Black	.013	-.011
	(.064)	(.023)
Masters degree	.143	.172
	(.268)	(.168)
Doctorate degree	.027	-.008
	(.032)	(.013)
Base Pay	351.642	230.880
	(593.253)	(325.709)
Total Pay	508.564	443.355
	(608.677)	(340.176)
Experience	-.334	.457
	(.831)	(.389)

Notes: The table presents the effect of APIP on the AP teacher characteristics. Results are presented in nominal terms. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

Figure 8: Number of Students per AP Classroom

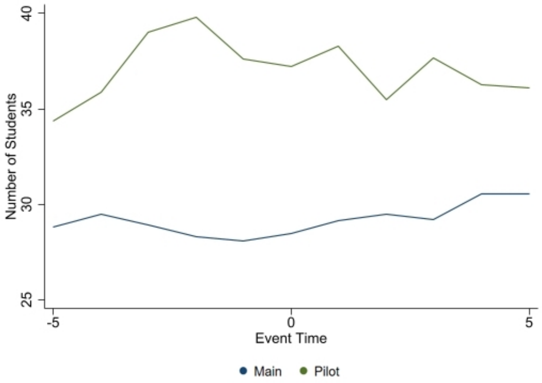


Table 11: Heterogeneity by AP Classroom Sizes

	Pilot		Scaled-up	
	<Median	>Median	<Median	>Median
Fund Per Student	1.743***	1.307***	2.394***	.386***
	(.464)	(.203)	(.680)	(.083)
Total AP Courses	.781**	1.970***	.012	.443
	(.380)	(.851)	(.231)	(.368)
AP Enrollment	-13.545	138.991**	-6.025	43.541**
	(8.681)	(64.845)	(5.279)	(20.730)

Notes: The table presents the effect of APIP on the number of AP courses and AP enrollment separated by treatment arms and their AP classroom sizes. Median is pre-treatment median percentage of AP classroom sizes across Texas. and, AP classroom size is defined by their position according to the median. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

only in schools with a higher AP classroom size. Moreover, the number of AP courses do not increase in scaled-up schools which has higher AP classroom sizes, supporting that the effect is not on the supply side. The table shows that AP classroom size is the most promising mechanism behind the difference between pilot schools and scaled-up schools.

Together with the heterogeneity in the number of AP courses offered in pre-period, the difference in AP classroom sizes provide evidence to think that the demand for AP classes is higher in the schools which we see positive effects of APIP.

7 Discussion

7.1 Access of Disadvantaged Students to AP Courses

I present evidence that indicates if students in a secondary school have high demand for AP courses, a funding dedicated for AP Program help increasing AP utilization. Nonetheless, the AP program is currently facing a persistent issue of underrepresentation of disadvantaged students who take advantage of it. Specifically, Black and Hispanic students enroll in AP courses and achieve lower grades at approximately half the rate of their White counterparts, with low income being the primary contributing factor (Klopfenstein 2004). Long et. al. (Long, Conger, and McGhee Jr 2019) argue that schools in less-resourced communities are unable to implement AP at the level expected by its founders. Descriptive evidence also reveals lower enrollment rates among Black, Hispanic, and economically disadvantaged students (Moore and Slate 2008).

Moreover, certain descriptive studies demonstrate that the AP program does not consistently provide the expected level of college readiness (Hallett and Venegas 2011) and fails to adequately motivate talented disadvantaged students (Kyburg, Hertberg-Davis, and Callahan 2007).

It might be speculated that these disparities in enrollment could potentially

Table 12: The Heterogeneity of Economic Status

	Pilot		Scaled-up	
	<Median	>Median	<Median	>Median
Total AP Courses	1.94***	1.56***	.32	-.07
	(.63)	(.54)	(.26)	(.19)
AP Enrollment	118.72***	39.66**	20.69*	-.86
	(44.85)	(18.11)	(11.98)	(11.79)

Notes: The table presents the effect of APIP on the number of AP courses and AP enrollment separated by treatment arms and their wealth status. Median is pre-treatment median percentage of free meal program enrollment across Texas. and, wealth status is defined by their position according to the median. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

be mitigated through policies aimed at promoting access to the AP program, with APIP being one such example. However, despite examining the differences concerning the economic status of schools and the composition of AP students, I could not find evidence supporting the notion that APIP helps increase AP utilization among disadvantaged students.

7.1.1 Access of Low Income Schools

Policies, even though they aim to help economically disadvantaged students, might not reach their goals. To see if there is a difference between the effect of APIP on economically advantaged and and economically disadvantaged schools, I create a variable, median percentage of students who are enrolled in free meal program. To create this variable, first I calculate percentage of students who are enrolled in free meal program each year for each school. Then I take the median of the percentage column. Thus, I end up having a median percentage of students who are enrolled in free meal program of all schools of all years. Then, I classify schools who have a higher percentage of students who are enrolled in free meal program than the median percentage of free meal students in Texas, as economically disadvantaged, and other schools as economically advantaged. As seen in the Table 12, for pilot schools, the increase in the AP enrollment after the implementation of APIP is three times higher in economically advantaged schools, even though the increase in the number of AP courses are similar. For scaled-up schools, I do not observe a change for economically disadvantaged schools, while there is an increase in AP enrollment for economically advantaged schools. These results suggest that even though the introduction of APIP aims to increase AP utilization by making enrollment accessible for students, it fails to reach economically disadvantaged schools.

Table 13: The Change in AP Student Composition

	Pilot	Scaled-up
At Risk	6.12	4.21
	(5.67)	(3.18)
Special Education	2.04**	-.13
	(.96)	(.39)
Black	7.31*	2.18*
	(4.05)	(1.23)
Hispanic	4.21	3.49
	(5.50)	(3.10)
Free meal	4.83	3.26
	(3.92)	(2.33)

Notes: The table presents the effect of APIP on the demographic composition of AP students separated by treatment arms. The results show the actual number of students. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.

7.1.2 The composition of AP students

Policies like APIP intend to target disadvantaged students. Although APIP does not have a component which specifically targets to racially disadvantaged students, the correlation between race and economic status makes us expect an improvement in AP utilization of racially and economically disadvantaged students. However, Table 13 present evidence that those disadvantaged students are not able to improve their AP utilization. AP enrollment for Students who are at risk, Black, Hispanic or who are enrolled in free meal program is not significantly increase.

When we consider the facts that schools which has lower rate of free meal enrollment shows an improvement and the enrollment for disadvantaged students are not increased, we can conclude that APIP might fail to address needs of disadvantaged students. Policy makers should consider specific needs of these students to close the AP utilization gap between them and advantaged students. One-on-one mentoring, college readiness educations and addressing pre-AP needs might be a couple of examples.

7.2 Cost - Benefit Analysis

This section uses results from Section 5 to provide a back-of-the-envelope calculation of costs and benefits of the APIP. These calculations help to understand the effectiveness of the program in pilot phase and in the scaled-up phase separately. I consider first five years of the pilot and scaled-up phases to calculate the total cost and benefits assuming the effect is constant after that period of time. The first reason behind this assumption is that, since the high school data stops at 2005, I can determine only the first five years of the program for the scaled-up schools, and to be in line with that I also consider only the first five years of the program in pilot schools. Second, the logarithmic nature of the outcomes make the assumption reasonable. The years included in pilot phase will be 1997 to 2001 while the years included in the scaled-up phase is 2001-2005. The total funding of the program for the first five years of pilot phase is $ 500,000 and of scaled-up phase is $10,000,000.

The results from Table 4 shows the first stage outcomes of the APIP. AP enrollment increased by 84 yearly in pilot schools and 57 of them are the first time enrollees. The number of schools which get APIP funding for the first time each year in the pilot phase is as follows: 16 schools in 1997, 7 schools in 1998, 33 school in 1999, and 31 schools in 2000. So the net increase of AP enrollment over the course of 1997-2001 is 22,596 with 15,333 new unique students while there is no statistically significant increase for scaled-up schools.

The results from Table 6 imply that graduating from a high school with APIP funding in the pilot phase increase earnings by about 7% yearly at the age of 25-30. I follow Chetty, Hendren, and Katz 2016 to predict the effect of this increase in income on lifetime earnings. I assume (i) the 7 percent increase in annual income is constant over the life cycle; (ii) the profile of income for students who graduated from high schools with APIP funding follows the US population average; (iii) the real wage growth rate is 0.5 percent; and (iv) the discount rate is 3 percent. Based on these assumptions, graduating from a high school with APIP funding in the pilot phase would increase pretax lifetime income by about $107,000 (present value of about $50,000).

Overall, the calculation suggests that getting APIP funding generates a high rate of return for pilot schools with the total expense is $ 500,000 and the total wage gain is $50,000 × 173 (mean number of high school graduates per pilot school). On the other hand, the $10,000,000 expense for the scaled-up schools do not generate a wage increase. The government's gain is calculated as $865,000 ($50,000 × 173 × .1) per school per year with the assumption of a ten percent increase in the tax revenue for the pilot phase and the loss is $7,000,000 for the scaled-up phase.

8 Conclusion

This paper uses the variation in the AP incentive funding caused by the staggered implementation of the Advanced Placement Incentive Program in Texas to quantify the effect of the program on AP program access and its long run outcomes. The program was introduced in 1997, aimed to make AP courses accessible for all students, and the schools implemented it after their application to the program. In 2000-2001 academic year, with the federal introduction of No Child Left Behind Program, the APIP was scaled up and the percentage of schools who have the program raised from 3% to 30 %. I use Callaway & Sant'anna method to estimate a difference-in-differences model under staggered implementation and dynamic heterogenous treatment effects. I find that implementing the APIP increases the number of AP courses by 2 courses, the AP enrollment by 84 students for the schools which implemented the program before scaling up and does not have a significant effect on the schools which implemented the program after scaling up. The heterogeneity carries over time for college outcomes and labor market outcomes.

I analyze possible mechanisms behind the heterogeneity. The most promising mechanism is the demand for AP courses. Although I cannot directly test the mechanism, I utilize AP classroom sizes to see if the students want to take AP courses if the courses are available. I find that the AP classroom sizes for pilot schools is higher than it is for scaled-up schools. The difference stays the same for both treatment arms over time, before and after the implementation of APIP, although the number of courses are increased for pilot schools. Moreover, scaled-up schools with higher number of students per AP classroom experience an increase in AP enrollment.

To conclude, APIP was successfully implemented in the pilot phase, while it failed to be scaled up. The most promising mechanism behind the failure of the scaling up is a demand-side mechanism, suggested by the difference in AP classroom sizes. With a simple back of the envelope calculation, in the first five years of the program for pilot schools the total APIP funding was $500,000 with an increase of approximately 35,000 in AP enrollment. In the first five years of the program for scaled-up schools, the total funding was $10M with no increase in the enrollment. Before scaling up, performing a demand analysis will be useful to prevent a failure. Moreover, if a policy can actually increase AP utilization, it has long run positive effects on college and labor market outcomes suggested by the reduced-form analysis.

References

Agency., Texas Education (2006). *Advanced Placement and International Baccalaureate examination results in Texas, 2004-05 (Document No. GE06 601 10).*

Andrabi, Tahir et al. (2020). “Upping the ante: The equilibrium effects of unconditional grants to private schools”. In: *American Economic Review* 110.10, pp. 3315–3349.

Athey, Susan and Guido W Imbens (2022). “Design-based analysis in difference-in-differences settings with staggered adoption”. In: *Journal of Econometrics* 226.1, pp. 62–79.

Banerjee, Abhijit et al. (2017). “From proof of concept to scalable policies: Challenges and solutions, with an application”. In: *Journal of Economic Perspectives* 31.4, pp. 73–102.

Banerjee, Abhijit V et al. (2007). “Remedying education: Evidence from two randomized experiments in India”. In: *The quarterly journal of economics* 122.3, pp. 1235–1264.

Barrera-Osorio, Felipe and Leigh L Linden (2009). “The use and misuse of computers in education: evidence from a randomized experiment in Colombia”. In: *World Bank Policy Research Working Paper* 4836.

Bold, Tessa et al. (2012). “Interventions & institutions experimental evidence on scaling up education reforms in Kenya”. In: *Preliminary draft. Available at http://www. iies. su. se/polopoly_fs/1.101632* 13481, p. 37980.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2022). “Revisiting Event Study Designs: Robust and Efficient Estimation”. In: *Available at SSRN 2826228*.

Callaway, Brantly and Pedro HC Sant’Anna (2021). “Difference-in-differences with multiple time periods”. In: *Journal of Econometrics* 225.2, pp. 200–230.

Chajewski, Michael, Krista D Mattern, and Emily J Shaw (2011). “Examining the role of Advanced Placement® exam participation in 4-year college enrollment”. In: *Educational Measurement: Issues and Practice* 30.4, pp. 16–27.

Chetty, Raj, Nathaniel Hendren, and Lawrence F Katz (2016). “The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment”. In: *American Economic Review* 106.4, pp. 855–902.

Code, Texas Education (1996). “Texas school law bulletin”. In: *Austin, TX: West Publishing.*

De Chaisemartin, Clément and Xavier d’Haultfoeuille (2020). “Two-way fixed effects estimators with heterogeneous treatment effects”. In: *American Economic Review* 110.9, pp. 2964–2996.

Duflo, Esther (2004). “Scaling up and evaluation”. In: *Annual World Bank Conference on Development Economics*. Vol. 2004.

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011). “Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya”. In: *American economic review* 101.5, pp. 1739–1774.

Ferguson, Ronald F (1991). “Paying for public education: New evidence on how and why money matters”. In: *Harv. J. on Legis.* 28, p. 465.

Fryer, Roland G, Steven D Levitt, John A List, et al. (2015). *Parental incentives and early childhood achievement: A field experiment in Chicago heights*. Tech. rep. National Bureau of Economic Research.

Gigliotti, Philip and Lucy C Sorensen (2018). “Educational resources and student achievement: Evidence from the Save Harmless provision in New York State”. In: *Economics of Education Review* 66, pp. 167–182.

Goodman-Bacon, Andrew (2021). “Difference-in-differences with variation in treatment timing”. In: *Journal of Econometrics* 225.2, pp. 254–277.

Hallett, Ronald E and Kristan M Venegas (2011). “Is increased access enough? Advanced placement courses, quality, and success in low-income urban schools”. In: *Journal for the Education of the Gifted* 34.3, pp. 468–487.

Hargrove, Linda, Donn Godin, and Barbara Dodd (2008). “College Outcomes Comparisons by AP® and Non-AP High School Experiences. Research Report No. 2008-3.” In: *College Board*.

Hong, Kai and Ron Zimmer (2016). “Does investing in school capital infrastructure improve student achievement?” In: *Economics of Education Review* 53, pp. 143–158.

Jackson, C Kirabo (2010). "A little now for a lot later a look at a texas advanced placement incentive program". In: *Journal of Human Resources* 45.3, pp. 591–639.

Jackson, C Kirabo and Claire L Mackevicius (n.d.). "What impacts can we expect from school spending policy? Evidence from evaluations in the US". In: *American Economic Journal: Applied Economics* ().

Klopfenstein, Kristin (2004). "Advanced placement: Do minorities have equal opportunity?" In: *Economics of Education Review* 23.2, pp. 115–131.

Klopfenstein, Kristin and M Kathleen Thomas (2009). "The link between advanced placement experience and early college success". In: *Southern Economic Journal* 75.3, pp. 873–891.

Krueger, Alan B (1999). "Experimental estimates of education production functions". In: *The quarterly journal of economics* 114.2, pp. 497–532.

Kumaranayake, Lilani (2008). "The economics of scaling up: cost estimation for HIV/AIDS interventions". In: *Aids* 22, S23–S33.

Kurowski, Christoph et al. (2007). "Scaling up priority health interventions in Tanzania: the human resources challenge". In: *Health policy and planning* 22.3, pp. 113–127.

Kyburg, Robin M, Holly Hertberg-Davis, and Carolyn M Callahan (2007). "Advanced Placement and International Baccalaureate programs: Optimal learning environments for talented minorities?" In: *Journal of advanced academics* 18.2, pp. 172–215.

List, John A (2022). *The Voltage Effect: HOW TO MAKE GOOD IDEAS GREAT AND GREAT IDEAS SCALE*. Currency.

Long, Mark C, Dylan Conger, and Raymond McGhee Jr (2019). "Life on the frontier of AP expansion: Can schools in less-resourced communities successfully implement advanced placement science courses?" In: *Educational Researcher* 48.6, pp. 356–368.

Mattern, Krista D, Jessica P Marini, and Emily J Shaw (2013). "Are AP® Students More Likely to Graduate from College on Time? Research Report 2013-5." In: *College Board*.

Mobarak, Ahmed Mushfiq (2022). *Assessing social aid: the scale-up process needs evidence, too.*

Monroe-DeVita, Maria, Gary Morse, and Gary R Bond (2012). "Program fidelity and beyond: multiple strategies and criteria for ensuring quality of assertive community treatment". In: *Psychiatric Services* 63.8, pp. 743–750.

Moore, George W and John R Slate (2008). "Who's taking the advanced placement courses and how are they doing: A statewide two-year study". In: *The High School Journal*, pp. 56–67.

Rothschild, Eric (1999). "Four decades of the advanced placement program". In: *The History Teacher* 32.2, pp. 175–206.

Smith, Jonathan, Michael Hurwitz, and Christopher Avery (2017). "Giving college credit where it is due: Advanced Placement exam scores and college outcomes". In: *Journal of Labor Economics* 35.1, pp. 67–147.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". In: *Journal of Econometrics* 225.2, pp. 175–199.

Theokas, C and R Saaris (n.d.). *Finding America's missing AP and IB students. shattering expectations series. Washington, DC: The Education Trust; 2013.*

Al-Ubaydli, Omar, John A List, and Dana L Suskind (2017). "What can we learn from experiments? Understanding the threats to the scalability of experimental results". In: *American Economic Review* 107.5, pp. 282–286.

A Tables and Figures

B Raw Data

The raw means of number of AP courses shows that there is an unexpected increase in the number of courses for scaled-up schools in 1999. The available resources such as TEA reports or Federal policy reports do not indicate any changes that may lead to the increase in the number of AP courses for scaled-up schools. This section tries to discover if the increase in the mean number of AP courses in scaled-up schools contaminate the null effect. Since Section 5 shows the positive effects for the pilot schools, I set a placebo treatment year of 1999 for scaled-up schools, to see this result is affecting other outcomes including the long run outcomes.

Figure 12 shows the results of the Callaway & Sant'Anna model for the outcomes of the number of AP courses, AP enrollment, College Enrollment and College Graduation where the treatment year is set to 1999 for all scaled-up schools. Unlike pilot schools, the positive and significant increase in the number of AP courses is not carried to other outcomes. These results help us to argue that the increase in the pre-period of scaled-up schools does not contaminate the causal null effect of APIP on the scaled-up schools.

C Density of Urbanicity Levels

D Labor Market Outcomes Unconditional on High School Graduation

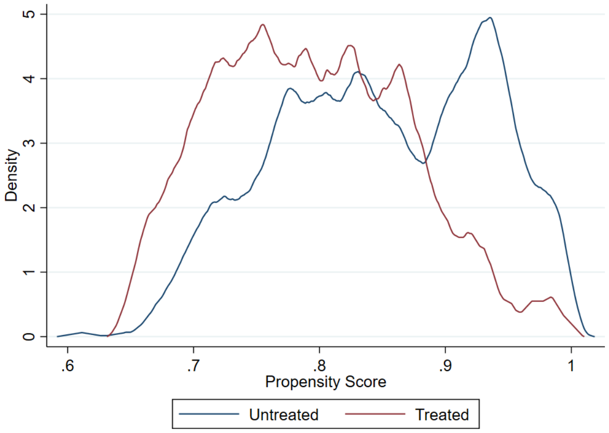


Figure 9: Pilot

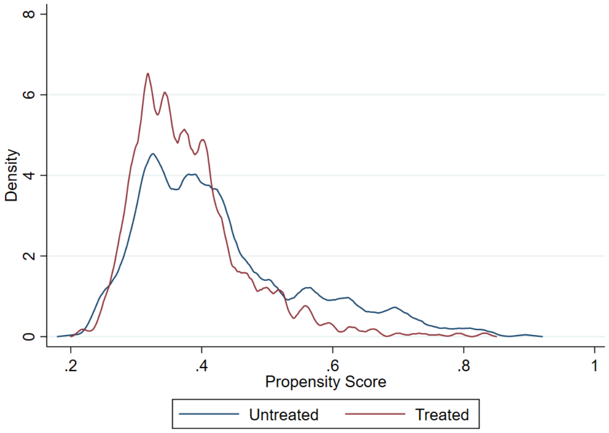
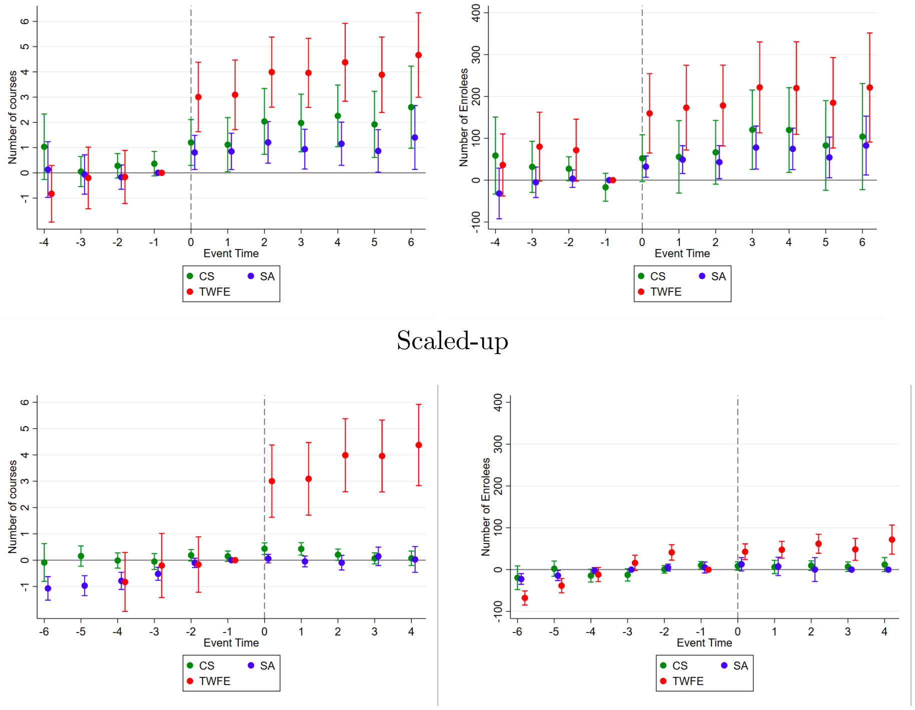


Figure 10: Scaled-up

Notes: These figures present the distribution of propensity scores of each treatment group and their comparison groups.

Figure 11: Sun & Abraham and TWFE results

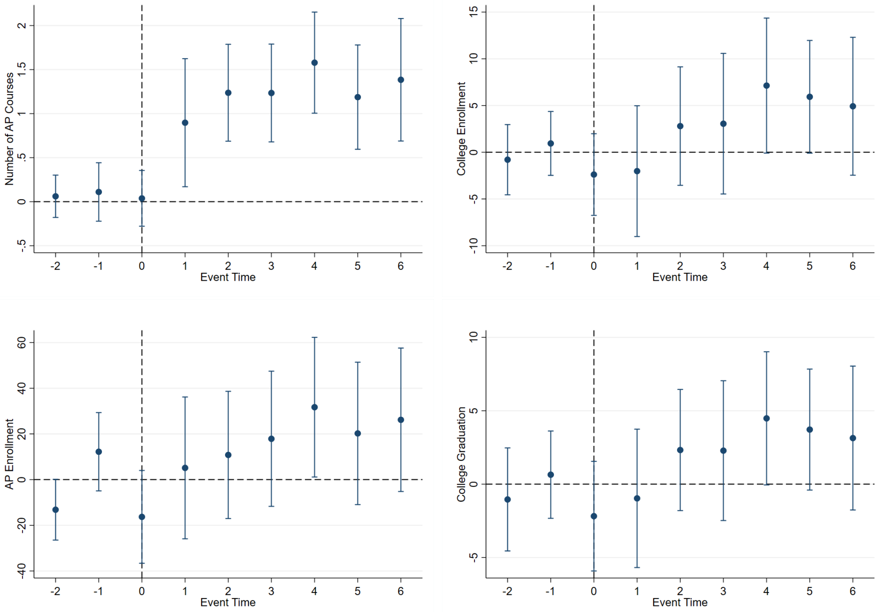


Pilot

Scaled-up

Notes: The figures present Sun & Abraham DiD and TWFE estimation results along with Callaway & Sant'Anna for robustness. The same covariates are used in all models. Standard errors are clustered at the district level.

Figure 12: Scaled-up schools - Treatment Year is 1999



Notes: The figures present raw means of the AP class sizes and the total number of AP courses side-by-side for each treated group of schools with respect to the implementation time of APIP.

Figure 13: Distribution of Schools with Respect to Urbanicity Levels

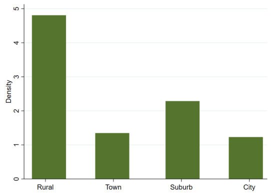


Figure 14: Pilot Schools

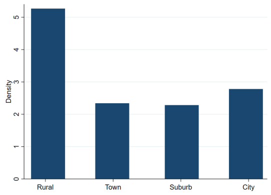


Figure 15: Scaled-up schools

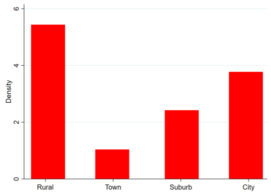


Figure 16: Never-Treated Schools

Notes: The figure presents the densities of pilot, scaled-up, and never-treated schools by their urbanicity level. The figures use raw data, unweighted to matching propensity scores.

Figure 17: Employment Rate and Wages

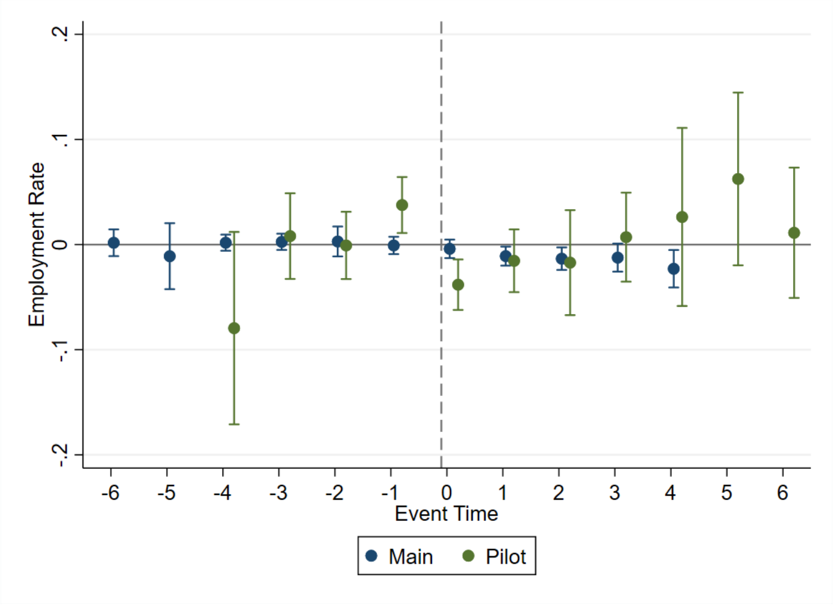


Figure 18: Employment Rate

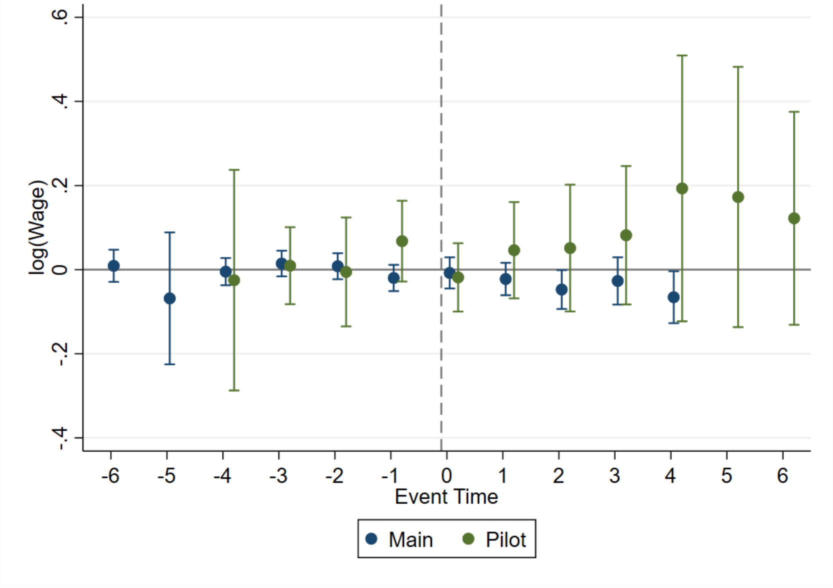


Figure 19: Wages

Notes: The figure presents the labor market outcomes of pilot, and scaled-up schools unconditional on high school graduation. School controls are included in all analyses and standard errors are clustered at district-level. Values in parentheses show standard errors.