ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-626

The Long-Run Impacts of Universal Pre-K: Evidence from the First Statewide Program

Jordan S. Berne
University of Michigan

Nearly all studies of preschool's long-run effects examine means-tested programs; little is known about the long-run effects of universal programs. A number of key differences—including population served, scale, and counterfactual options—may cause universal programs to have different effects than previously studied means-tested programs. Using a difference-in-differences framework, I estimate the effects of Georgia's first-in-the-nation statewide universal pre-K program on adult educational attainment and employment. The program made children 4.5 percent more likely to graduate from high school and 13.7 percent more likely to obtain a bachelor's degree (although the latter effect is imprecise). I find similar results in a supplemental analysis that uses the synthetic control method. I find no effects on associate degree attainment or employment.

VERSION: August 2022

Suggested citation: Berne, Jordan S.. (2022). The Long-Run Impacts of Universal Pre-K: Evidence from the First Statewide Program. (EdWorkingPaper: 22-626). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/k4bh-0114

The Long-Run Impacts of Universal Pre-K: Evidence from the First Statewide Program*

Jordan S. Berne†
University of Michigan

August 2022
Preliminary and incomplete

Abstract

Nearly all studies of preschool's long-run effects examine means-tested programs; little is known about the long-run effects of *universal* programs. A number of key differences—including population served, scale, and counterfactual options—may cause universal programs to have different effects than previously studied means-tested programs. Using a difference-in-differences framework, I estimate the effects of Georgia's first-in-the-nation statewide universal pre-K program on adult educational attainment and employment. The program made children 4.5 percent more likely to graduate from high school and 13.7 percent more likely to obtain a bachelor's degree (although the latter effect is imprecise). I find similar results in a supplemental analysis that uses the synthetic control method. I find no effects on associate degree attainment or employment.

Keywords: Georgia, universal preschool, state-funded pre-K, long-run effects, educational attainment, labor market outcomes

*I would like to thank my third-year faculty readers, Michael Mueller-Smith and Christina Weiland, and many others who offered useful comments, including but not limited to Bruno Ferman, Ariza Gusti, Brian Jacob, Michael Ricks, and Nathan Sotherland.

†Email: jsberne@umich.edu

Contents

1 Introduction

In the last 60 years, there has been a dramatic expansion of public preschool in the United States. The expansion began in 1965 with the creation of the federal Head Start program. It has continued at the state level, with more than 40 states funding preschool programs today.[1] This movement has been driven, in part, by the belief that early childhood education can improve life trajectories well into adulthood—a belief supported by extensive evidence from the academic literature. Researchers have shown that preschool can increase educational attainment and employment, reduce crime, and improve health outcomes.

The most influential results on preschool's long-run effects come from means-tested programs that began in the 1960s and 70s: the Perry Preschool Project, the Carolina Abecedarian Project, and Head Start. Since then, one of the major changes in the public preschool landscape has been the emergence of universal programs, i.e., programs in which eligibility does not depend on family income. The first statewide universal pre-K (UPK) program was established in Georgia in 1995. Today, eight states, the District of Columbia, and cities such as Boston, Chicago, New York, and San Antonio all have some form of universal pre-K. At the federal level, one of the marquee proposals in President Biden's domestic agenda is the creation of a $200 billion federal UPK program for three- and four-year-olds.

Nearly all prior studies on preschool's long-run effects examine means-tested programs—very few have examined universal programs. This lack of research is not due to a lack of interest; rather, until recently, universal programs were too young to *have* long-run effects. In this paper, I help fill this gap in the literature by exploiting the relatively long existence of Georgia's first-in-the-nation statewide UPK program. I estimate its impact on adult educational attainment and employment.

My primary analysis uses a difference-in-differences (DiD) framework that leverages variation in exposure to Georgia UPK among people born in different states and years. The variation in exposure comes from the introduction of the program in 1995. To assess robustness, I also conduct a secondary analysis that uses the synthetic control method (SCM). In both analyses, I use data from the American Community Survey. I find evidence across both methods that UPK

[1] Only Idaho, Indiana, Montana, New Hampshire, South Dakota, and Wyoming do not have state-funded preschool programs (Friedman-Krauss et al., 2021).

increases educational attainment as an adult. Children *exposed* to UPK are 1.7 percentage points more likely to graduate high school and 1.8 percentage points more likely to obtain a bachelor's degree (although the latter effect is not always statistically significant). After scaling the intent-to-treat effects by program enrollment, children who *participate* in UPK are 4.0 percentage points (4.5 percent) more likely to graduate high school and 4.3 percentage points (13.7 percent) more likely to obtain a bachelor's degree. On the other hand, I find no effects on associate degree attainment or employment.

My findings advance our knowledge of preschool's long-run effects in a few ways. First, as already mentioned, very little is known about the long-run effects of *universal* preschool. There are at least two theoretical reasons why the effects might differ from means-tested programs: 1) mixing children from high- and low-income families may generate different peer effects, and 2) children from families with high incomes have higher-quality outside options than children targeted by means-tested programs. In short, universal programs serve a different population, so they might have different long-run effects.

Second, this paper improves our understanding of the long-run effectiveness of *large-scale* preschool programs, whether universal or not. Education policy has a long history of difficulty maintaining quality in large-scale programs (Murnane and Willett, 2011). Although Head Start is a major counterexample, Perry Preschool and Carolina Abecedarian were both small-scale and intensive to an extent that is challenging to replicate on a large scale. This casts some doubt on our ability to extrapolate their long-run impacts to statewide programs like Georgia UPK. As states and the federal government contemplate new preschool programs, it is important to accumulate evidence across a variety of large-scale programs.

Third, this paper comments on the long-run effectiveness of public preschool amidst a relatively modern early childhood environment. In many contexts today, public preschool has several close substitutes, which makes children's counterfactual experiences particularly important. As alternative programs proliferate, we should expect the marginal impact of each new one to diminish. Relative to the early childhood landscape in the 1960s and 70s—when the previously mentioned programs were introduced—the counterfactual programs in the early years of Georgia UPK bear a much greater resemblance to today's.

My findings indicate that Georgia UPK's long-run impacts are comparable to estimates from other preschool programs in the literature. As expected, Perry

Preschool's and Carolina Abecedarian's impacts on high school graduation are toward the top of the distribution. Georgia UPK's impact is a little smaller than Boston UPK's and similar to Head Start's. Overall, this suggests that large-scale UPK programs can have meaningful long-run impacts, even amidst relatively modern counterfactual environments.

The rest of this paper proceeds as follows. Section 2 provides background on Georgia's UPK program. Section 3 discusses the related preschool literature. Section 4 outlines my empirical strategy. I present results in Section 5 and compare them to the literature in Section 6. Section 7 concludes.

2 Universal Pre-K in Georgia

2.1 Program History and Enrollment

Georgia's pre-K program did not begin as statewide or universal. In 1993, with funding from a new lottery game, the Georgia state government created a voluntary, means-tested pre-K program to combat poor statewide educational outcomes. The program started small, operating in only 20 sites. Then, after two years of piloting, the state government re-established the program in 1995, expanding it across the state and opening eligibility to all children who turned four by September 1st. In doing so, Georgia pre-K became the first statewide UPK program in the country.

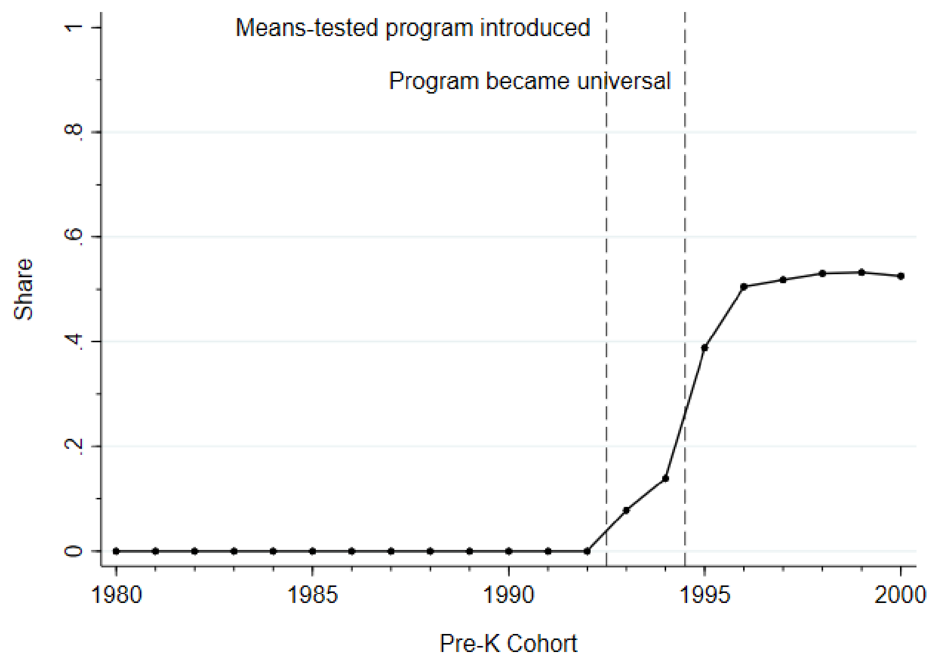
Rather than operate UPK classrooms directly, the Georgia state government provided funding to local providers that met UPK program requirements. Around 90 percent of all UPK classrooms were run by public school systems and private for-profit centers, but local providers also included public and private non-profits and Head Start centers (Bryan and Henry, 1998; Henry et al., 2003a, 2004).[2] To receive UPK funding, providers had to offer services for at least 6.5 hours a day during the local school year; classrooms could have no more than 20 students and had to maintain a minimum staff-to-child ratio of 1:10; teachers and classroom assistants had to have certain credentials; and providers had to use a pre-approved curriculum. Nearly two-thirds of all classrooms used the HighScope curriculum

[2] In the 1996-97 school year, among 203 randomly selected classrooms across the state, 47 percent of all providers were private for-profit, 43 percent were local school systems, 5 percent were Head Start Centers, 4 percent were private non-profit, and 2 percent were public non-profit (Henry et al., 2003a).

(Bryan and Henry, 1998).

As Figure 1 illustrates, enrollment in Georgia pre-K grew quickly. In the 1993-94 school year, when the means-tested program began, 8 percent of all four-year-olds in the state enrolled. By 1996-97, the second year of the universal program, enrollment reached 51 percent. Of course, raw enrollment does not represent a causal increase in preschool participation. Using a regression discontinuity approach, Fitzpatrick (2010) estimates that UPK caused a 14 percent increase in preschool enrollment of any type in 2000. The greatest increases were in less densely populated rural areas. For families with higher income, UPK partially crowded out private preschool enrollment (Cascio and Schanzenbach, 2013).

Figure 1. Share of Four-Year-Olds Enrolled in Georgia's Pre-K Programs



Notes: Participation numbers are taken from Fitzpatrick (2010), who draws from Brackett et al. (1999), various web sources, and the Census Bureau's Time Series of State Population Estimates by Age.

To my knowledge, administrative records on the universe of children in Georgia UPK do not exist for the earliest years of the program. However, we can get a sense of the children's demographics from studies that surveyed classrooms across the state. In a random sample from the 1996-97 school year, Henry et al. (2003a) find that around 50 percent of enrolled children were white, non-Hispanic and around 39 percent were Black, non-Hispanic. Around half of enrolled children were eligible for free or reduced-price lunch, and around 25 percent had health insurance coverage through Medicaid or Georgia's SCHIP program. Roughly 70 percent of enrolled children lived in a household with married parents or a parent and a significant other.

2.2 Program Quality

Georgia UPK was mediocre to good on "structural quality," which refers to program inputs like classroom size and per-child spending. As previously mentioned, the state government guaranteed some structural quality by linking funding to program inputs. The National Institute of Early Education Research reported that Georgia UPK met 7 of 10 structural quality standards in the 2001-02 school year, nearly 2 points higher than the average for comparable state pre-K programs (Barnett et al., 2003). However, among a random sample of Georgia UPK classrooms in the same school year, trained assessors found that the quality of space, furnishings, and activities were not quite "good," on average (Henry et al., 2003b).

In terms of "process quality," which refers to the social, emotional, physical, and instructional aspects of children's day-to-day experiences, Georgia UPK was good but not great. Process quality likely influences child outcomes more directly than structural quality, although the two are related. Many of the early studies on Georgia UPK used the Early Childhood Environment Rating Scale (ECERS) to assess program quality. Within ECERS, there are two subscales that most closely approximate process quality: "interactions" and "language-reasoning." The interactions subscale measures the quality of student-student interactions, teacher-student interactions, and general supervision. The language-reasoning subscale measures how well students are encouraged to communicate, develop reasoning skills, and engage with books and other materials. ECERS scores range from 1 to 7, with 1=inadequate, 3=minimal, 5=good, and 7=excellent. Among a random sample of classrooms in the 2001-02 school year, Georgia UPK received a mean score of 5.5 for interactions and 4.9 for language-reasoning. Both scores were higher (but not statistically different) than the analogous scores for Head Start classrooms in Georgia (Henry et al., 2003b).

Overall, the quality of Georgia UPK may be described as mediocre to good. Combining all (structural and process) ECERS subscales, the average program quality of Georgia UPK (among a random sample of classrooms) in the 1997-98 school year was 4.66. Average scores were similar across region and provider type (Bryan and Henry, 1998). Although program quality was not great in an absolute sense, it was slightly better than average among alternative options. Tietze et al. (1996) calculate a mean ECERS score of 4.26 across 401 preschool classrooms in four states in the 1992-93 school year. In a sample of 32 Head Start classrooms in the South, Bryant et al. (1994) calculate a mean ECERS score of 4.24 in the 1990-91 and 1991-92 school years.

2.3 Short- and Medium-Run Impacts

Prior research on the early years of Georgia UPK finds that it had positive effects on short- and medium-run academic outcomes.[3] Using a propensity score matching strategy and data from the 2001 pre-K cohort, Henry et al. (2006) find that UPK was more effective than Head Start for children from families with low income. At the beginning of kindergarten, UPK students outperformed Head Start students on cognitive assessments and teacher evaluations. Fitzpatrick (2008) and Cascio and Schanzenbach (2013) investigate medium-run test score impacts using data from the National Assessment of Educational Progress. Both studies use the same difference-in-differences framework (around the introduction of the program) that I use. Fitzpatrick (2008) finds that UPK exposure improved fourth grade math and reading scores by 6 to 12 percent of a standard deviation for school-lunch-eligible children in rural areas. It also reduced the probability of grade retention. Cascio and Schanzenbach (2013) find suggestive evidence that UPK improved eighth grade math scores for students from families with low income.[4] Although it's not uncommon in the preschool literature for effects to fade out and then reemerge, I argue that detectable effects in the medium run lend credibility to the long-run effects I uncover.

As with any early childhood program, one of the possible mechanisms for improved child outcomes is maternal labor supply. However, there is no strong evidence that the introduction of UPK raised employment among Georgian mothers

[3]Studies on Georgia UPK in more recent years have more mixed findings. A recent study using a lottery admissions design finds that Georgia UPK had slightly negative impacts on fourth grade math and reading test scores for children who qualify for free or reduced-price meals (Woodyard et al., 2022). Importantly, though, these findings are not closely related to the impacts I estimate; Woodyard et al. (2022) examine a single district in Metro Atlanta and pre-K cohorts 14 to 20 years younger than the youngest cohort I examine.

[4]The effect on test scores obtained by Cascio and Schanzenbach (2013) is a weighted average from Georgia's and Oklahoma's UPK programs.

(Fitzpatrick, 2010; Cascio and Schanzenbach, 2013). Another possible mechanism is that UPK might have lowered costs for families who otherwise would have paid for private childcare. Indeed, Cascio and Schanzenbach (2013) show that families spent less on childcare after the introduction of UPK.

3 Related Literature

The earliest evidence on the long-run effects of preschool comes from two small-scale early childhood interventions: the Perry Preschool Project (PPP) and Carolina Abecedarian (ABC). PPP operated from 1962 to 1967, and ABC operated from 1972 to 1982. The programs share several features: researchers designed and implemented them, assignment to treatment was randomized, and they almost exclusively served Black children from families with low income. Both programs were highly intensive. PPP included four home visits a month and provided parenting instruction. ABC provided health and nutrition care, childcare goods, and counseling. Program participants—who have now been followed into their 30s and 40s—have more schooling, higher employment rates, and fewer arrests than their control group counterparts (Heckman et al., 2010; Belfield et al., 2006; Elango et al., 2016; Campbell et al., 2012). PPP and ABC helped establish that preschool can have long-lasting impacts, but these programs differ on a number of dimensions from programs like Georgia UPK. Large-scale programs today are less intensive and serve more diverse children.

Another program that has received substantial research attention, and that is more similar to Georgia UPK, is Head Start. Head Start is a federal, means-tested pre-K program established in 1965. The Head Start literature generally finds positive long-run effects on educational attainment and income (Bailey et al., 2021; Owen, 2018; Johnson and Jackson, 2019; Garces et al., 2002; Deming, 2009). Bailey et al. (2021) find that Head Start increased high school completion by 2.7 percent and college completion by 39 percent for participants in the program's early years. Notably, however, Pages et al. (2020) find that Head Start's long-run impacts are smaller—possibly even negative—for more recent cohorts, relative to home care. The heterogeneity across cohorts may be due to household conditions improving across cohorts. Compared to PPP and ABC, Head Start operates in a much wider range of communities, has a more complex administrative structure, and spends less per child. Georgia UPK is more comparable to Head Start than

PPP and ABC on these dimensions. A key difference between Georgia UPK and Head Start, though, is universal eligibility.

Although universal early childhood programs are much newer than means-tested programs, there are two in North America whose long-run effects have been evaluated. The first program, established in 1997, is a universal childcare program for kids ages zero to four in Quebec, Canada. Strikingly, Baker et al. (2008) find that the program had negative contemporaneous effects on noncognitive and health outcomes. In a follow-up paper, Baker et al. (2019) show that the negative effects persisted in the long run; as adults, participants had lower life satisfaction, worse health, and higher crime rates. These results are surprising, but may be partly explained by treatment effect heterogeneity that the overall results conceal. Kottelenberg and Lehrer (2017) find that Quebec's program improved some developmental outcomes for more disadvantaged children, but that these effects were offset by negative effects on more advantaged children. This result highlights the important role outside options play in mediating the effects of public pre-K.

The second program—and the one more similar to Georgia UPK—is UPK in Boston. Gray-Lobe et al. (2021) study the effects of Boston UPK on children who participated between 1997 and 2003 and find positive effects on long-run educational outcomes. Boston UPK participants were 9 percent more likely to graduate high school and 18 percent more likely to enroll in college "on-time." These effects were primarily concentrated among boys, with little heterogeneity by race or family income. One mechanism appears to have been improved disciplinary outcomes in high school. Though similar in many respects, my study differs from the Boston study in the following key ways: 1) I examine a statewide program that operates in a wider range of communities and facility types; 2) I observe adult outcomes beyond educational attainment;[5] 3) the Boston study explores a large set of mechanisms throughout K-12 schooling, whereas my data do not allow me to; and 4) I use a quasi-experimental DiD approach, whereas the Boston study exploits a random admissions lottery. The Boston research design almost certainly has greater internal validity, but its externally validity is likely more limited. Within Boston, the results are valid for the 48 percent of pre-K students who attended oversubscribed programs. It's not clear whether the results would hold

[5] In the next iteration of this paper, I plan to estimate effects on additional adult outcomes, such as income, receipt of government benefits, and crime.

for the other 52 percent of students who attended non-oversubscribed programs, much less students from areas very different than Boston. By studying children from all across Georgia who enrolled in all sorts of classrooms, my estimates may be more externally valid for other states considering UPK programs.

Though the literature on UPK's long-run effects is quite small, there is a larger literature on UPK's short- and medium-run effects. This literature mostly finds that UPK improves cognitive skills and educational outcomes (Phillips et al., 2017). Outside of Georgia, researchers have found positive effects of UPK in Oklahoma and Boston using age-eligibility regression discontinuity designs (Gormley Jr. and Gayer, 2005; Weiland and Yoshikawa, 2013). In an examination of Boston UPK in the 2008-09 school year, Weiland and Yoshikawa (2013) find positive short-run effects on language, literacy, math, socioemotional skills, and executive functioning. The literature is not uniformly positive though. In a meta-analysis of universal early childhood education programs in North America, Europe, and Australia, van Huizen and Plantenga (2018) find that 34 percent of all estimated effects are significantly positive and 16 percent are significantly negative. A couple recent papers have directly compared universal and means-tested statewide pre-K programs. Cascio (2021) and Zerpa (2021) find that both types of pre-K generally have positive effects on early educational outcomes for children from families with low income, but that effects are larger for universal programs. The present study helps determine whether these patterns persist in the long run.

4 Empirical Strategy

4.1 Data and Sample Construction

My analysis uses annual survey data from 2005 through 2020 from the U.S. Census Bureau's American Community Survey (ACS). The ACS has information on an extensive set of adult outcomes. In this iteration of the paper, I focus on educational attainment and employment. The measures of educational attainment I examine are high school graduation (including GEDs), associate degree attainment, and bachelor's degree attainment. Because the data is cross-sectional, I cannot link adults to their childhood characteristics, which limits my pre-treatment controls to time-invariant characteristics like race, ethnicity, sex, and quarter of birth.

Another limitation from not observing adults as children is that individual enrollment in UPK is unobserved. However, I can approximate exposure to UPK by sorting individuals into (state of birth) × (year of birth + 4) pre-K cohorts.[6] All adults born in Georgia beginning with the 1995 pre-K cohort are considered exposed to UPK. As an example, consider an individual observed at the age of 30 in the 2014 ACS. This person would be in the 1988 pre-K cohort because they would have been four years old in that year. Regardless of their state of birth, they would not have been exposed to UPK. Now consider another individual, also observed in the 2014 ACS, but at age 20. This person would be in the 1998 pre-K cohort. If they were born in Georgia, they would have been exposed to UPK.[7]

My definition of UPK exposure might not perfectly capture the population of interest since it is based on state of birth rather than state of residence at age four. It has the advantage, though, of avoiding bias from families moving to Georgia because of UPK. In practice, the two definitions would likely produce similar results. Only 14 percent of all children born in Georgia leave the state by age four.[8] In Section 5.1.3, I calculate suggestive treatment effects that account for this out-migration from Georgia.

After sorting individuals into pre-K cohorts, I restrict my sample to those old enough to have realized long-run outcomes. Ideally, I would observe everyone once their adult outcomes were fully determined. In practice, someone may appear in the ACS years before they finish their education, for example. One solution would be to set a minimum age across the board for inclusion in the sample. However, there is a trade-off in the age threshold choice: older ages mean greater outcome stability but smaller sample sizes. To maximize the sample for each outcome, I choose different age thresholds based on visual inspection of Appendix Figure A1. I attempt to choose the youngest age at which each outcome appears highly stable. In my primary specification, I use 20 as the minimum age for high

[6] I define "year of birth" to align with academic cohorts, not calendar years. The birthday cutoff for kindergarten in Georgia is September 1st. Therefore, each academic cohort consists of individuals born in the fall of calendar year t and in the winter, spring, and summer of calendar year $t+1$. This is the 12-month grouping I use to define "year of birth," and I label the whole period as year t.

[7] For a full description of the observed age range and sample size of each pre-K cohort, see Appendix Table A1.

[8] I calculate out-migration from Georgia using a 1-in-20 national sample from the 2000 Census.

school graduation, 22 for associate degree attainment, 24 for bachelor's degree attainment, and 24 for employment. Because stabilization ages are subjective, I conduct a sensitivity analysis in Section 5.3.1.

Table 1. Summary Statistics for the 1987-1998 Pre-K Cohorts

	Overall	Control Group States	Georgia
Race and ethnicity (%)			
White	78.6	80.6	66.6
Black	10.9	7.9	29.5
Other	10.4	11.5	4.0
Hispanic	5.8	6.3	2.5
Sex (%)			
Female	49.5	49.5	49.5
Long-run outcomes (%)			
Graduated high school	91.4	91.9	88.5
Obtained associate degree	10.4	10.8	7.8
Obtained bachelor's degree	33.0	33.3	31.2
Employed	75.9	76.5	72.5
Observations	864,009	741,905	122,104

Notes: Author's calculations using American Community Survey data. Following Zerpa (2021), the 18 control group states are Alabama, Alaska, Arizona, Hawaii, Idaho, Indiana, Iowa, Minnesota, Mississippi, Montana, Nevada, New Hampshire, North Dakota, Rhode Island, South Dakota, Utah, Washington, and Wyoming. The minimum age threshold for the number of observations and the race, ethnicity, and sex shares is 20. For the long-run outcomes, the minimum age thresholds are 20 for high school graduation, 22 for associate degree attainment, 24 for bachelor's degree attainment, and 24 for employment. The race categories "white," "black," and "other" are mutually exclusive; when their shares do not sum to 100 it is due to rounding.

For a description of the ACS sample, see Table 1.[9] People born in Georgia are more likely to be Black than people born in control group states. They are less likely to be white, of another race, or Hispanic. As adults, people born in Georgia have lower educational attainment, on average, and are less likely to be employed. Recall that poor statewide educational outcomes was one of the motivations for the creation of Georgia UPK.

[9]Table 1 provides summary statistics for the 1987-1998 pre-K cohorts, which aligns with the difference-in-differences analysis. The synthetic control analysis uses data on the 1980-1998 pre-K cohorts. The summary statistics for these cohorts are very similar.

4.2 Research Design

4.2.1 Primary Analysis: Difference-in-Differences

In my primary analysis, I estimate UPK's long-run impacts using a difference-in-differences (DiD) approach. The first "difference" is a comparison of people born in Georgia who turned four before and after the introduction of UPK. The second "difference" is a comparison of people born in Georgia (i.e., the treatment group) and people born in other states (i.e., the control group). Since the treatment is UPK exposure, this framework estimates intent-to-treat (ITT) effects, which are highly policy-relevant for a statewide preschool program. With so many close substitutes, UPK's impact crucially depends on whether children are induced into the program, not just its effects on the treated.

Letting Y_{isc} denote an adult outcome for individual i from state of birth s in pre-K cohort c, my primary model specification is:

$$Y_{isc} = \beta_0 + \beta_1 UPK_i + \beta_2 MTPK_i + \theta \mathbf{X}_i + \alpha_s + \gamma_c + \epsilon_{isc}, \tag{1}$$

where UPK_i is a binary indicator for being exposed to Georgia UPK, $MTPK_i$ is a binary indicator for being exposed to Georgia's means-tested pre-K program, α_s is a vector of state of birth fixed effects, and γ_c is a vector of pre-K cohort fixed effects. $\mathbf{X}_i$ is a vector that contains a cubic age polynomial and time-invariant pre-treatment controls (race, ethnicity, sex, and quarter of birth). The ITT effect of interest is β_1.[10]

I estimate Equation 1 using data on pre-K cohorts 1987 through 1998. This sample includes six Georgia cohorts exposed to neither of Georgia's pre-K programs, two exposed to Georgia's means-tested pre-K program, and four exposed to Georgia UPK.

Following Zerpa (2021), I limit my control group to the 18 states that had no more than a very small statewide pre-K program during the time of my analysis.[11] See Appendix Figure A2 for a map of the these states. This restriction is not strictly necessary, but it yields a cleaner comparison group and interpretation of

[10]Note that the issues sometimes associated with two-way fixed effects models are not present in my setting because treatment adoption is not staggered, all control units are never treated, and treatment never turns off once it's on.

[11]The 18 states identified by Zerpa (2021) are Alabama, Alaska, Arizona, Hawaii, Idaho, Indiana, Iowa, Minnesota, Mississippi, Montana, Nevada, New Hampshire, North Dakota, Rhode Island, South Dakota, Utah, Washington, and Wyoming.

β_1.

Inference in this setting is tricky because my sample has too few states for reliable use of most asymptotic methods. With only 19 clusters (i.e., states), the standard cluster-robust variance estimator would not perform well. Moreover, some inference methods that perform well with a small number of clusters, like the Wild cluster bootstrap, would not perform well here because there is only one treated state (i.e., Georgia). Instead, I use the Ferman and Pinto (2019) method, which is designed for situations like this. This method is an extension of the cluster residual bootstrap and the Conley and Taber (2011) method. It uses information from control states to approximate variation in the treated state, allowing for arbitrary correlation within state. Following Ferman and Pinto (2019), I model heteroskedasticity generated by variation in state size, and I allow for state sizes to change across cohorts.

4.2.2 Secondary Analysis: Synthetic Control Method

I also conduct a secondary analysis that uses the synthetic control method (SCM). This analysis is useful for investigating situations where there are concerns about parallel trends. It also provides a natural setting to conduct permutation inference.

SCM constructs a counterfactual for Georgia by taking a weighted average of ”donor pool” states (Abadie et al., 2010). The weights are chosen so that the weighted average, called “synthetic Georgia,” closely approximates the real Georgia in the pre-treatment period. Assuming synthetic Georgia provides an accurate counterfactual, differences in outcomes in the post-treatment period can be interpreted as UPK’s causal effects.

To implement the SCM analysis, I use pre-K cohorts from 1980 through 1998 since SCM is more credible with a longer pre-treatment period. I use the 18 states identified by Zerpa (2021) to form the donor pool.

My setting differs from the typical synthetic control setting in that I begin with microdata rather than aggregate data. Therefore, I aggregate the ACS data to the (state of birth) $\times$ (year of birth + 4) level. Before doing so, though, I purge the outcomes of differences due to observable individual characteristics by estimating the following model on data from pre-K cohorts 1980-1992:

$$Y_{is} = \beta_0 + \theta \mathbf{X}_i + \alpha_s + u_{is}, \tag{2}$$

where the vector $\mathbf{X}_i$ contains the same covariates as in Equation 1. Then, I use the estimated parameters to calculate residuals for the entire sample, which I use as the outcome when aggregating. This procedure makes Georgia and the donor pool states more similar and should therefore improve the match between Georgia and synthetic Georgia.

After aggregating the data, I use a variation of the traditional SCM estimator to calculate UPK's effects. The variation is that the SCM estimator is applied to data that has been demeaned using each state's pre-treatment mean.[12] Demeaning forces SCM to match on trends rather than levels, as with traditional DiD (Ferman and Pinto, 2021; Doudchenko and Imbens, 2016). This is important in my setting because Georgia's educational outcomes tend to be at the bottom of the donor pool distribution. Letting $\tau_{s,c}$ denote the difference between real and synthetic outcomes for cohort c of state s, I calculate Georgia UPK's ITT effect as $\frac{1}{4}\sum_{c=1995}^{1998} \tau_{Georgia,c}$.

Following the recent SCM literature, my primary specification uses every pre-treatment outcome as a predictor and no other predictors (Ferman and Pinto, 2021; Doudchenko and Imbens, 2016). I check for robustness in Section 5.3.2 by estimating models that include additional predictors and only a few lagged outcomes.

I conduct inference in the SCM analysis using a permutation method, as proposed by Abadie et al. (2010). First, I apply the SCM estimator separately for Georgia and each donor pool state using the procedure described above. Next, for each state, I calculate the root mean square error (RMSE) between real and synthetic outcomes in the pre- and post-treatment periods, i.e., $RMSE_{s,pre} \equiv \sqrt{\frac{1}{13}\sum_{c=1980}^{1992} \tau_{s,c}^2}$ and $RMSE_{s,post} \equiv \sqrt{\frac{1}{4}\sum_{c=1995}^{1998} \tau_{s,c}^2}$.[13] These measures capture how similar the real and synthetic outcomes are. Finally, for each state, I compute a test statistic: $T_s \equiv RMSE_{s,post}/RMSE_{s,pre}$. Intuitively, T_s should be relatively large for Georgia (if UPK has a nonzero treatment effect) because $RMSE_{s,post}$ should be relatively small in placebo states (since they had no treatment). Dividing by $RMSE_{s,pre}$ accounts for states that have large post-treatment "effects" because of a bad fit with their synthetic counterparts. UPK's long-run effects are statistically significant if $T_{Georgia}$ is extreme relative to the donor pool T_s distribution. This

[12] I demean the data via the state of birth fixed effects in Equation 2.

[13] Notice that I do not include the 1993 and 1994 cohorts in either measure, as discussed in Section 4.2.3.

exercise is essentially a placebo falsification test.

The permutation-based inference in the SCM analysis complements the regression-based inference in the primary analysis. In the DiD analysis, inference quantifies uncertainty in the parameter estimate due to sampling variability. In the SCM analysis, inference quantifies uncertainty in the parameter estimate due to the possibility of chance effects. These are different thought experiments. Conducting inference both ways strengthens the causal interpretation of any estimated effects.

4.2.3 Identifying Assumptions

To interpret the estimated effects as causal, three assumptions must be met. First, both analyses require some version of a ”parallel trends” assumption. In the DiD analysis, the assumption is that the treatment and control groups would have followed parallel trends in the absence of UPK. I provide evidence supporting this assumption in Section 5.1.1 using dynamic event study models. The analogous SCM assumption is that synthetic Georgia correctly shows how outcomes would have evolved in real Georgia in the absence of UPK. I assess this assumption by examining how well synthetic Georgia matches real Georgia in the pre-treatment period.

The second assumption is that there were no effects on pre-K cohorts that weren't exposed to treatment (i.e., the ”no anticipation” assumption). Such effects would contaminate the counterfactual and bias the post-treatment effects. In this context, there may be effects on the 1993 and 1994 pre-K cohorts stemming from Georgia's short-lived means-tested pre-K program. Fortunately, this type of potential violation is straightforward to handle. In the DiD analysis, I estimate separate effects for the cohorts exposed to the means-tested program. In the SCM analysis, I exclude the 1993 and 1994 pre-K cohorts when calculating UPK's treatment effects and the T_s test statistics.

The third and final assumption needed for identification is that no other policies or shocks occurred in Georgia (or control states) in between UPK exposure and the measurement of adult outcomes. Such an occurrence would conflate the effects of UPK with the effects of the policies/shocks. To my knowledge, there are no policies/shocks large enough to threaten a causal interpretation. Notably, Cascio and Schanzenbach (2013) make the same assumption to estimate the impact of UPK on eighth-grade test scores, although the risk of conflation is smaller in their case because they examine medium-run outcomes.

5 Results

5.1 Difference-in-Differences Results

5.1.1 Pre- and Post-Treatment Effect Dynamics

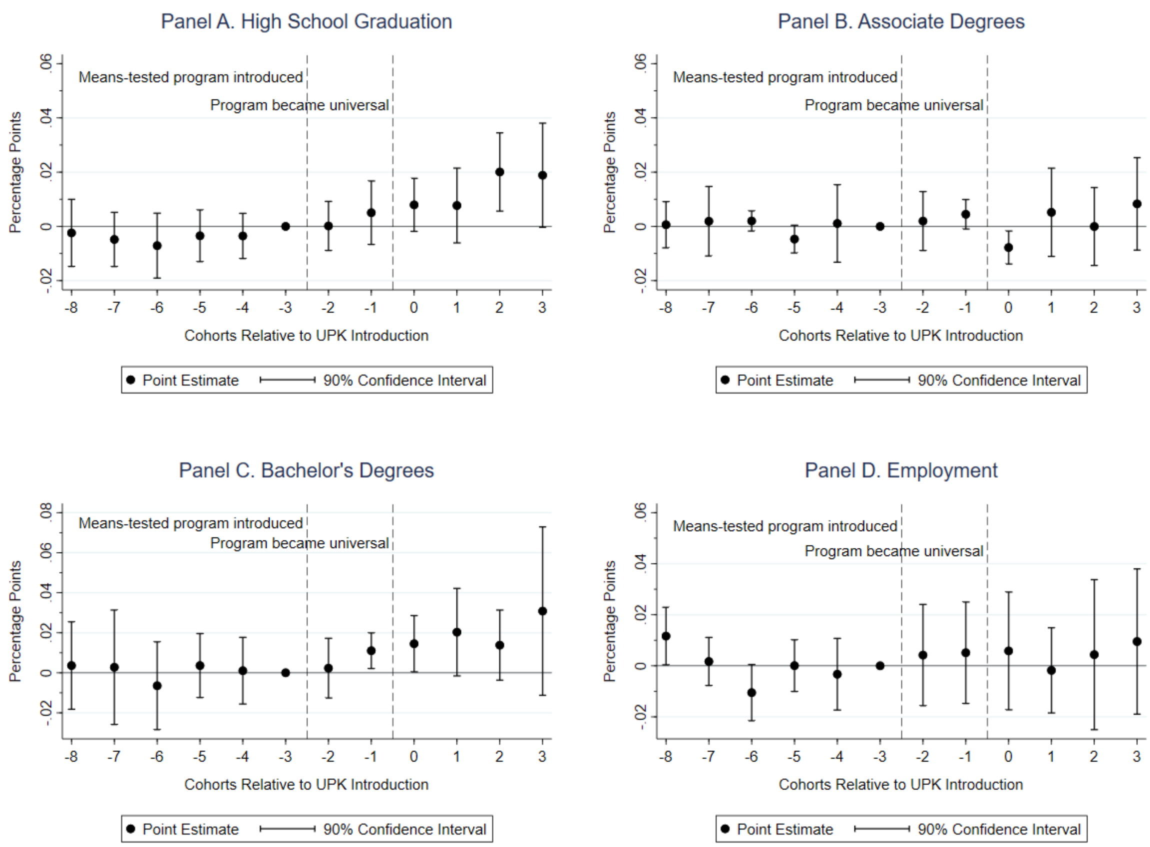
Before discussing the main DiD estimates, Figure 2 shows the results from flexible event study models that are useful for evaluating the parallel trends assumption.[14] The effects for cohorts -8 through -3 show how Georgia was trending relative to the control group before piloting the means-tested pre-K program. These effects provide the cleanest test of the parallel trends assumption. The effects for cohorts -2 and -1 are not quite as clean, but enrollment in the means-tested pre-K program was small enough that they may still be informative, especially for cohort -2.

The case for parallel trends looks particularly strong for associate and bachelor's degree attainment. Nearly all of the pre-treatment coefficients for these outcomes are small and statistically indistinguishable from 0.

The pre-treatment trends for high school graduation and employment are somewhat convincing, but not completely. For high school graduation, what appears to be a positive treatment effect may be a continuation of an upward trend that began among untreated cohorts. This pre-treatment trend would be small, but it merits some concern. The synthetic control analysis addresses this concern by constructing a counterfactual with the same pre-treatment trends. For employment, the pre-treatment effects for cohorts -8 through -6 are not ideal, but the effects for the five cohorts leading up to treatment are more convincing.

[14] The Ferman and Pinto (2019) inference method is designed for DiD models, not event studies. However, an event study is essentially a collection of multiple 2×2 DiD estimates. Therefore, to apply the inference method to an event study, I estimate a separate 2×2 DiD model for each event time coefficient.

Figure 2. Event Study Plots of UPK's Long-Run Impacts



Notes: Author's calculations from estimating an event study version of Equation 1 using American Community Survey data. Inference is conducted using the Ferman and Pinto (2019) method, as described in Section 4. Cohort 0 refers to the 1995 pre-K cohort. The coefficients for cohorts -8 through -3 are the pre-treatment effects of interest. The coefficients for cohorts 0 through 3 are UPK's treatment effects.

Besides evaluating the parallel trends assumption, Figure 2 is also useful for investigating post-treatment effect dynamics. The effects on high school graduation and bachelor's degree attainment both rise gradually across treated cohorts. Notably, Cascio and Schanzenbach (2013) find the same pattern of rising treatment effects on eighth grade test scores. This consistency is reassuring since we would expect rising medium-run effects to translate into rising long-run effects. That they find the same pattern using different data and for a different outcome is further reassuring. One partial explanation for this pattern is that there was a marked increase in UPK enrollment after the first year. It's also possible that the program became more effective over time, as was the case with Oklahoma UPK (Hill et al., 2015).

A somewhat surprising result in the event studies (for high school graduation and bachelor's degrees) is the similarity in effect size for cohorts -1 and 0. The positive effects for cohort -1 are not surprising on their own, but 14 percent of cohort -1 enrolled in Georgia's means-tested pre-K program whereas 39 percent of cohort 0 enrolled in UPK. Based on enrollment alone, one might expect larger ITT effects for cohort 0. A couple details may partially explain this phenomenon. For one, the means-tested program may have had especially large effects because it targeted children with great need. Second, as just mentioned, research on Oklahoma UPK has found that large-scale programs can take time to become effective (Hill et al., 2015).

5.1.2 Main DiD Results

Having found mostly supportive evidence for the parallel trends assumption, the primary DiD results are presented in Table 2. The strongest result is for high school graduation. The estimate implies that UPK exposure made children 1.7 percentage points more likely to graduate high school. Relative to Georgia's pre-treatment high school graduation rate, which was 88 percent, this is a 1.9 percent increase.

UPK exposure also made children 1.8 percentage points more likely to obtain a bachelor's degree, but this effect is estimated less precisely. Off of a pre-treatment base of 31.3 percent, this would be a 5.8 percent increase. Although this effect is not statistically significant in the DiD analysis, it is in the SCM analysis using the permutation-based inference method. Overall, I consider the effect on bachelor's degree attainment suggestive.

Table 2. DiD Estimates of UPK's Long-Run Impacts

	High School Graduation	Associate Degrees	Bachelor's Degrees	Employment
Universal pre-K	0.017**	0.000	0.018	0.003
(standard error)	(0.009)	(0.010)	(0.015)	(0.010)
(p-value)	(0.044)	(0.988)	(0.295)	(0.738)
Pre-treatment mean	0.880	0.080	0.313	0.720
Minimum age	20	22	24	24
Observations	864,009	740,672	606,393	606,393
R^2	0.033	0.010	0.054	0.039

*** p<0.01, ** p<0.05, * p<0.1

Notes: Author's calculations from estimating Equation 1 using American Community Survey data. Inference is conducted using the Ferman and Pinto (2019) method, as described in Section 4.2.1. The "pre-treatment means" are the average adult outcomes of those in the 1987-1992 pre-K cohorts born in Georgia.

Conversely, UPK exposure has no detectable effect on associate degree attainment. The estimate is an imprecise 0.0. Interpreting this estimate is tricky though because the overall effect might mask offsetting sub-effects. On one hand, UPK might make those on the margin of going to any college more likely to go to community college. This would be a positive effect on associate degree attainment. On the other hand, UPK might cause some who would have gone to community college to go to four-year college instead, which would be a negative effect on associate degree attainment. Because these effects work in opposite directions, it is not clear how to interpret the overall effect.

There are no detectable effects on being employed as an adult either. The point estimate is an imprecise 0.3 percentage points.

5.1.3 Effects on the Treated

Recall that all of the results in Table 2 are ITT effects. Policymakers may also be interested in the effects of UPK *on the treated* (i.e., ATT effects). Here I provide suggestive estimates of the ATT effects by scaling the ITT effects to account for UPK enrollment and out-migration from Georgia. These ATT effects would be causal if UPK had no spillovers, but it almost certainly must have. Still, this back-of-the-envelope exercise is useful for getting a ballpark sense of the ATT effects.

Among all children born in Georgia, roughly 86 percent still reside there at age four. In the first four years of UPK, roughly 49 percent of all four-year-olds enrolled. Combining these figures, approximately $.86 * .49 \approx .42$ of exposed individuals in my sample enrolled in Georgia UPK. The ATT effect of Georgia UPK on high school graduation is therefore $1.7/.42 \approx 4.0$ percentage points (4.5 percent). The ATT effect on bachelor's degree attainment is $1.8/.42 \approx 4.3$ percentage points (13.7 percent), which is particularly large.

5.2 Synthetic Control Method Results

This section presents the results from the secondary SCM analysis. Reassuringly, the SCM results generally concur with the DiD results. Figures 3 and 4 compare trends between Georgia and synthetic Georgia for each outcome. Table 3 shows, for each outcome, the weight placed on each donor pool state and the quality of synthetic Georgia's pre-treatment fit. Figure 5 shows the placebo effect distributions used to conduct inference.

Table 3. SCM Donor Pool Weights and Pre-Treatment Period Fit, by Outcome

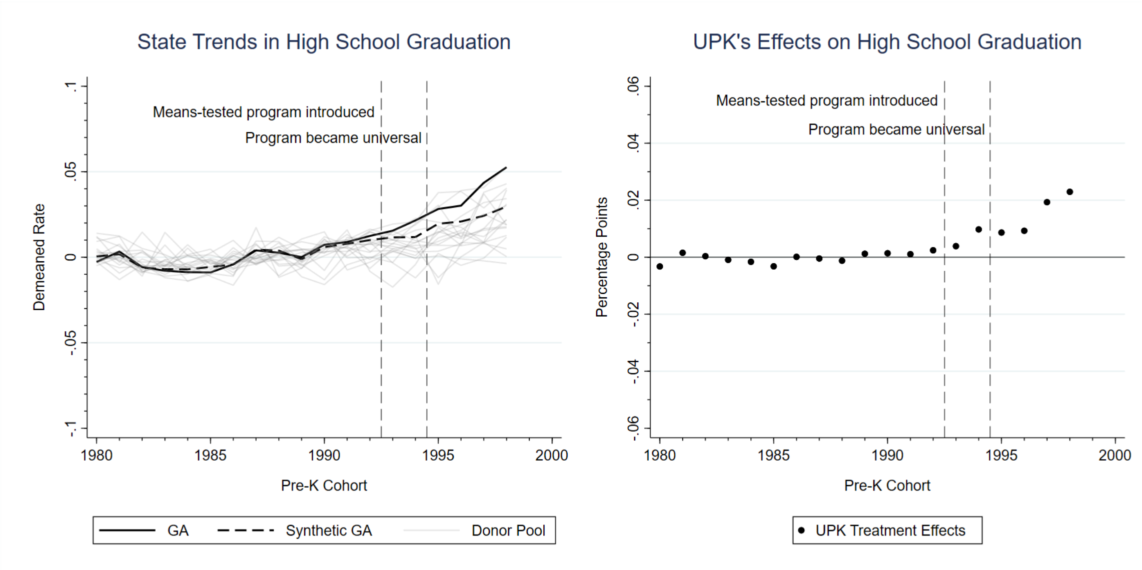
	High School Graduation	Associate Degrees	Bachelor's Degrees	Employment
Donor Pool Weights				
Alabama	0.023	0	0	0
Alaska	0	0.071	0	0.050
Arizona	0	0	0	0
Hawaii	0.342	0	0	0.024
Idaho	0	0.029	0	0.065
Indiana	0	0	0.104	0
Iowa	0.122	0.111	0	0.496
Minnesota	0.129	0.064	0.298	0
Mississippi	0.187	0.187	0.124	0.093
Montana	0	0	0	0
Nevada	0.112	0.165	0.182	0
New Hampshire	0.015	0.056	0	0
North Dakota	0	0.138	0.079	0.026
Rhode Island	0.070	0	0.087	0
South Dakota	0	0	0.126	0
Utah	0	0	0	0
Washington	0	0.179	0	0
Wyoming	0	0	0	0.246
Pre-Treatment Period Fit				
RMSE	0.0017	0.0003	0.0030	0.0034

Notes: Author's calculations using American Community Survey data on pre-K cohorts 1980 through 1992.

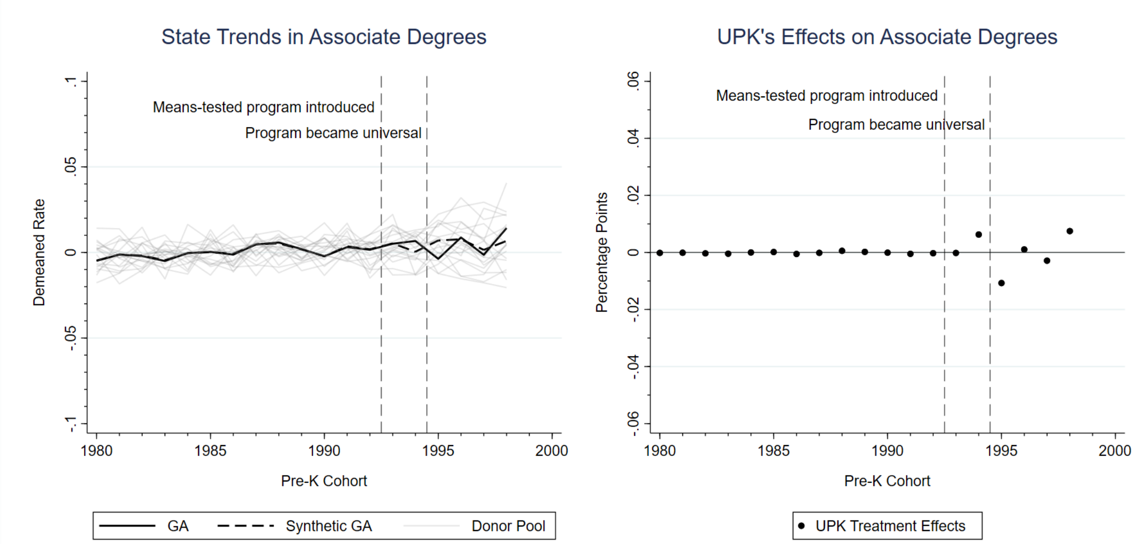
Beginning with high school graduation, Panel A of Figure 3 shows that the SCM estimator produces a synthetic Georgia that very closely resembles Georgia in the pre-treatment period. This helps alleviate concerns that the DiD estimate is driven by nonparallel trends. Most of the weight in the donor pool is placed on Hawaii, Iowa, Minnesota, Mississippi, and Nevada. The average ITT effect of UPK on high school graduation is 1.5 percentage points, which is very similar to the DiD estimate of 1.7 percentage points. Panel A of Figure 5 shows that the effect is highly statistically significant; $T_{Georgia}$ is greater than all 18 donor pool test statistics.

Figure 3. SCM Plots of UPK's Long-Run Impacts I

Panel A. High School Graduation



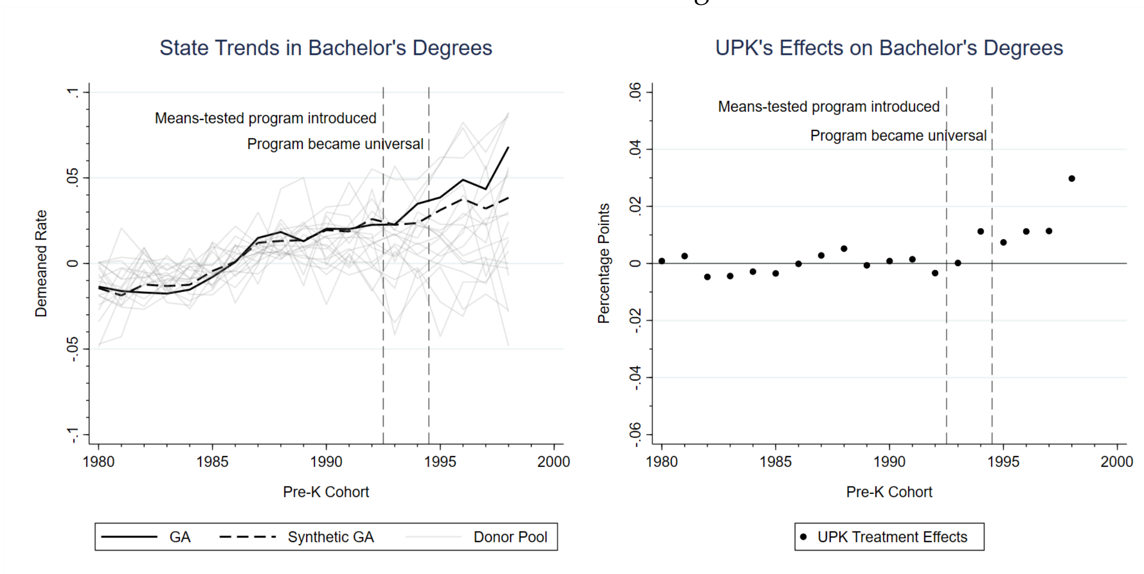
Panel B. Associate Degrees



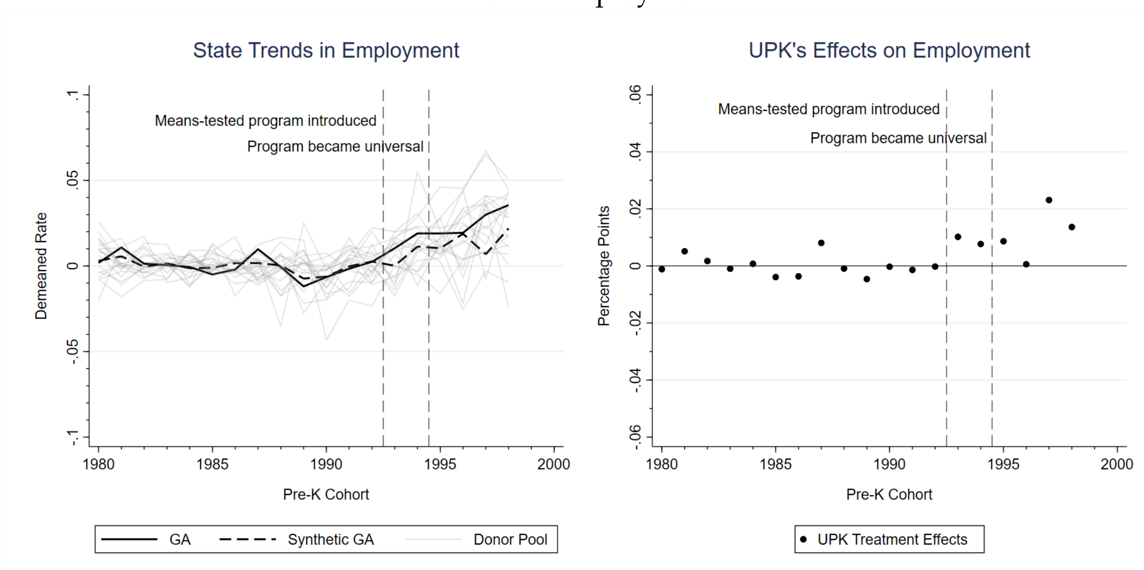
Notes: Author's calculations using American Community Survey data. In the plots on the right, each point is a simple difference between the outcomes of Georgia and synthetic Georgia in the plots on the left.

Figure 4. SCM Plots of UPK's Long-Run Impacts II

Panel A. Bachelor's Degrees



Panel B. Employment



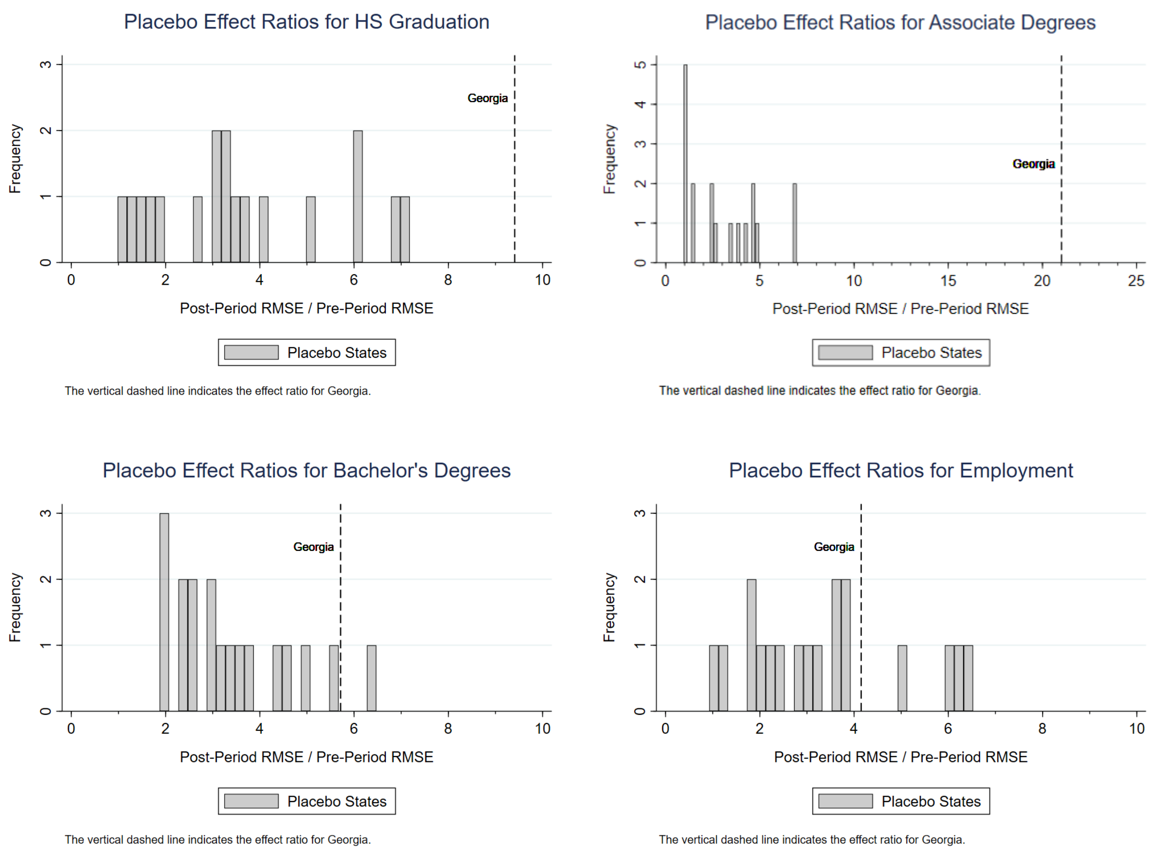
Notes: Author's calculations using American Community Survey data. In the plots on the right, each point is a simple difference between the outcomes of Georgia and synthetic Georgia in the plots on the left.

The SCM results for associate degree attainment are consistent with the DiD results too. Panel B of Figure 3 shows that synthetic Georgia has a remarkably good fit in the pre-treatment period. Georgia's pre-treatment RMSE is 5.5 times smaller than the smallest donor pool RMSE and 15.4 times smaller than the average donor pool RMSE.[15] Most of the weight in the donor pool is placed on

[15]Georgia's pre-treatment RMSE for associate degree attainment is also 5.7 times smaller than

Iowa, Mississippi, Nevada, North Dakota, and Washington. The average ITT effect of UPK on associate degree attainment is −0.1 percentage points, which is very similar to the DiD estimate of 0.0 percentage points. Inference for the effect on associate degree attainment is complicated by Georgia's exceptionally small pre-treatment RMSE, which inflates $T_{Georgia}$. Even with a very small effect size, $T_{Georgia}$ is far greater than every donor pool test statistic (as shown in Panel B of Figure 5). But, when $T_{Georgia}$ is recalculated using the smallest donor pool pre-treatment RMSE, it's greater than only 11/18 donor pool test statistics, which is consistent with the effect being statistically insignificant.

Figure 5. SCM Test Statistic (T_s) Distributions



Notes: Author's calculations using American Community Survey data. Test statistics are computed as described in Section 4.2.2.

The SCM results for bachelor's degree attainment are mostly consistent with the DiD results. Panel A of Figure 4 shows that synthetic Georgia has a fairly good fit in the pre-treatment period. The fit is not quite as good as for the previous

its pre-treatment RMSE for high school graduation, which is Georgia's next smallest. See Table 3.

two outcomes, but the pre-treatment differences between Georgia and synthetic Georgia are small relative to the post-treatment differences. Most of the weight in the donor pool is placed on Indiana, Minnesota, Mississippi, Nevada, and South Dakota. The average ITT effect of UPK on bachelor's degree attainment is 1.5 percentage points, which is slightly smaller but still very similar to the DiD estimate of 1.8 percentage points.

Contrary to the DiD result, UPK's effect on bachelor's degree attainment is statistically significant in the SCM analysis. As Panel C of Figure 5 shows, $T_{Georgia}$ is greater than 17/18 donor pool test statistics. Because the effect is statistically significant in the SCM analysis but not the DiD analysis, I view it as suggestive.

On employment, the SCM analysis finds some evidence of a positive effect, whereas the DiD analyses does not. Panel B of Figure 4 shows that synthetic Georgia has a fairly good fit in the pre-treatment period—similar to the fit in the bachelor's degree analysis. Most of the weight in the donor pool is placed on Iowa and Wyoming. The average ITT effect of UPK on employment is 1.1 percentage points, compared to 0.3 percentage points in the DiD analysis. As Panel D of Figure 5 shows, $T_{Georgia}$ is larger than 14/18 donor pool test statistics, which is not statistically significant at conventional levels. Overall, the SCM analysis provides only suggestive evidence of a positive effect on employment. In conjunction with the DiD analysis, the evidence for an effect of UPK on employment is weak.

5.3 Robustness

To probe the validity of my results, this section presents robustness checks against two research design questions that lack obvious answers. The first question is who to exclude from the sample because they are too young for their long-run outcomes to have stabilized. The second question is how to specify the SCM estimator.

UPK's impacts on associate degrees, bachelor's degrees, and employment are largely robust. There are some robustness concerns regarding the impact on high school graduation, but they are not great enough to invalidate a causal interpretation.

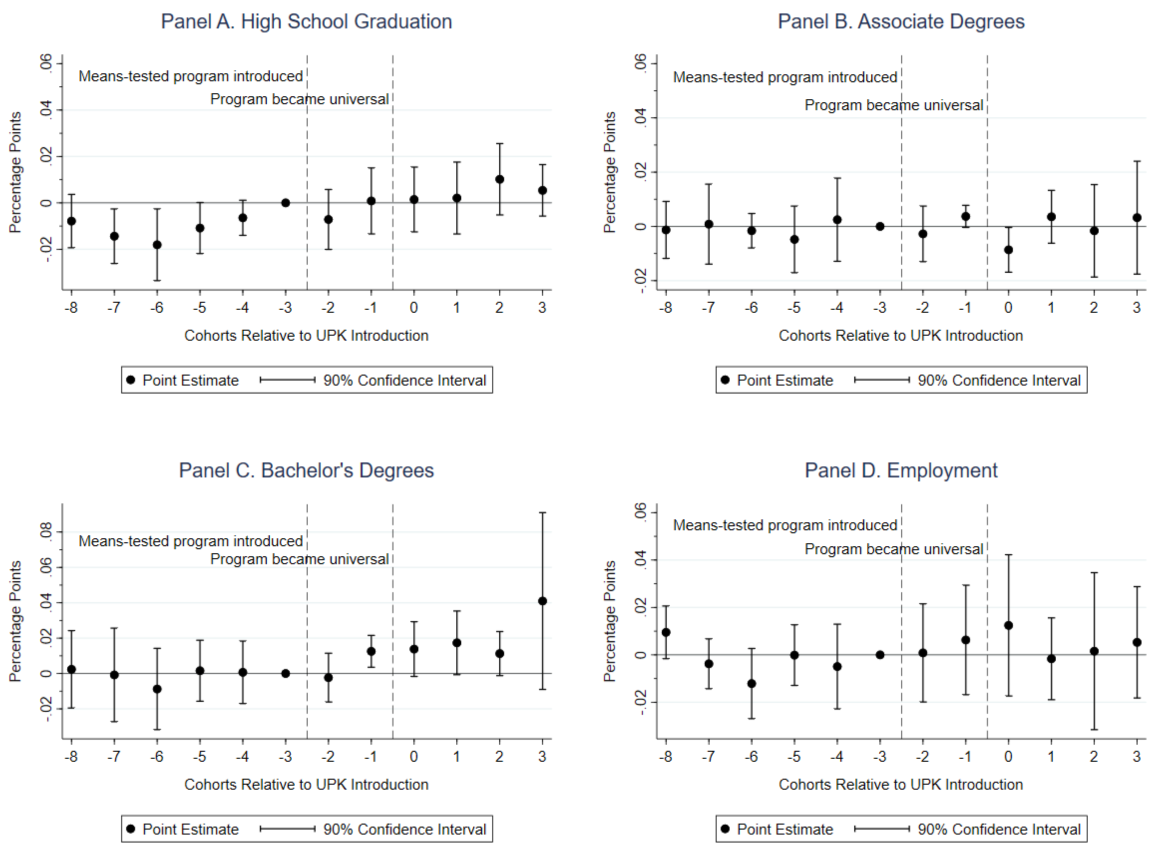
5.3.1 Alternative Minimum Age Threshold

Imposing a minimum age for inclusion in the sample requires an active decision by the researcher. In my primary specification, to maximize the sample for each outcome, I use 20 for high school graduation, 22 for associate degree attainment, 24 for bachelor's degree attainment, and 24 for employment. In this section, I test the sensitivity of my primary results by imposing a minimum age threshold of 25 for all outcomes. This threshold follows Bailey et al. (2021), who use the same basic research design to examine the effects of Head Start on similar adult outcomes.

The point estimates I obtain using 25 as the minimum age are highly similar to my original estimates. Re-estimating Equation 1 with the new sample produces ITT effects of 1.4 percentage points for high school graduation (compared to 1.7), -0.1 for associate degree attainment (compared to 0.0), 1.9 for bachelor's degree attainment (compared to 1.8), and 0.6 for employment (compared to 0.3). As before, the effect on high school graduation is statistically significant and the other effects are not.

For three of the four outcomes, the case for the parallel trends assumption also appears unchanged with the new sample. As Figure 6 illustrates, the pre-treatment trends for associate degree attainment, bachelor's degree attainment, and employment are essentially the same as with the primary samples.

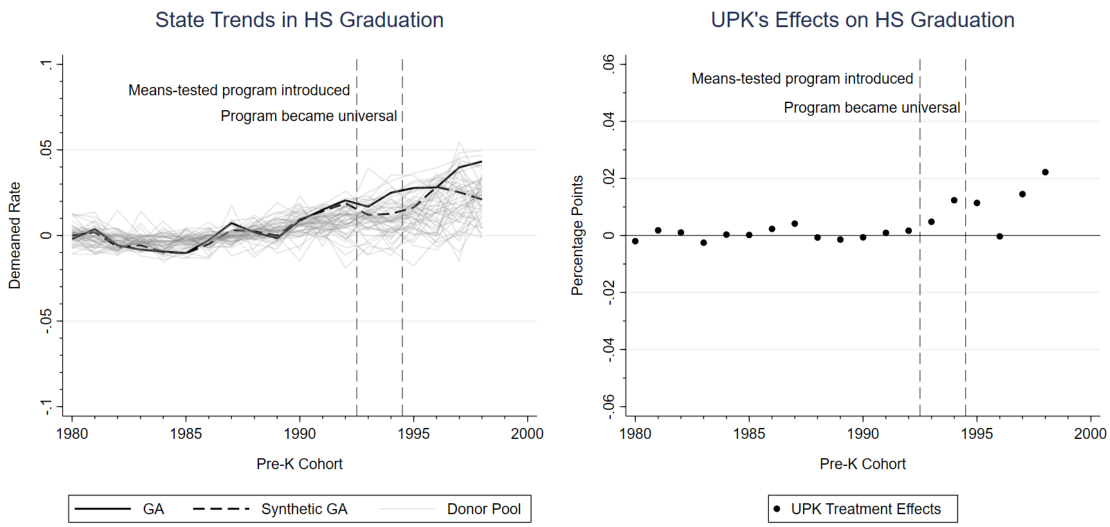
Figure 6. Event Study Plots of UPK's Long-Run Impacts, with 25 as the Minimum Age Threshold



Notes: Author's calculations from estimating an event study version of Equation 1 using American Community Survey data. For each outcome, the minimum age threshold for inclusion in the sample is 25. Cohort 0 refers to the 1995 pre-K cohort. The coefficients for cohorts -8 through -3 are the pre-treatment effects of interest. The coefficients for cohorts 0 through 3 are UPK's treatment effects.

Conversely, the case for parallel trends for high school graduation is concerning in Figure 6. The pre-treatment trends follow the same general pattern as with the primary sample, but they are more pronounced. These trends are driven by a sharp drop in high school graduation for Georgia's 1987-1989 pre-K cohorts followed by a sharp increase for the 1989-1992 cohorts. I privilege the primary results because they come from a larger sample with smoother trends. Nevertheless, I conduct an SCM analysis with the 25+ sample to check whether the results hold compared to a synthetic Georgia that exhibits the same sharp pre-treatment trends. None of the 18 donor pool states in my primary sample have trends as sharp as Georgia for the 1987-1992 pre-K cohorts, so I broaden the donor pool to include all other states.[16] The results are presented in Figure 7. Reassuringly, they are highly similar to the DiD and SCM results that come from the primary sample. I therefore view the effects on high school graduation as robust to the alternative minimum age threshold.

Figure 7. SCM Plots of UPK's Long-Run Impacts on High School Graduation, with 25 as the Minimum Age Threshold



Notes: Author's calculations using American Community Survey data. The minimum age threshold for inclusion in the sample is 25, and the donor pool includes all states other than Georgia. In the plot on the right, each point is a simple difference between the outcomes of Georgia and synthetic Georgia in the plot on the left.

5.3.2 Alternative SCM Specifications

In this section, I test the sensitivity of the results to alternative SCM specifications by varying the estimator along two dimensions. The first dimension is whether or not the outcomes are residualized using Equation 2 before the microdata is aggregated. The second dimension is the set of state-level predictors used to create the synthetic control. In some specifications, I use all lagged outcomes and no other predictors. In other specifications, I use additional state-level predictors and only a few lagged outcomes (those of the 1982, 1987, and 1992 pre-K cohorts). The additional predictors are the share of four-year-olds in school, the share of K-12 students in private school, the shares of mothers and fathers with children less than five years old who have completed 12th grade, the shares of mothers

[16]With the exception of broadening the donor pool, all other aspects of this SCM analysis are as described in Section 4.2.2.

Table 4. Robustness to Alternative SCM Specifications

	Specification Characteristics			Long-Run ITT Effects			
Estimator	Residualized Outcomes	All Lagged Outcomes	Additional Predictors	High School Graduation	Associate Degrees	Bachelor's Degrees	Employment
Primary estimator	✓	✓	✗	1.5	−0.1	1.5	1.1
				(18/18)	(18/18)	(17/18)	(14/18)
Alternative estimator I	✗	✓	✗	1.4	0.0	2.2	1.4
				(18/18)	(18/18)	(17/18)	(15/18)
Alternative estimator II	✓	✗	✓	0.9	−0.3	1.8	1.6
				(18/18)	(18/18)	(15/18)	(14/18)
Alternative estimator III	✗	✗	✓	0.8	−0.4	1.8	1.3
				(17/18)	(18/18)	(16/18)	(11/18)

Notes: Author's calculations using American Community Survey data. For each estimator, the first row contains the SCM point estimates and the second row contains the fraction of donor pool test statistics smaller than Georgia's. Larger fractions indicate greater statistical significance. The "Residualized Outcomes" column indicates whether the ACS microdata is residualized using Equation 2 before being aggregated. Note that regardless of whether I residualize the outcomes, I always demean them at the state-of-birth level before aggregating. The "All Lagged Outcomes" column indicates whether the outcomes for every pre-K cohort between 1980 and 1992 are used as predictors. The "Additional Predictors" column indicates whether predictors other than lagged outcomes are used. When additional predictors are used, the estimator uses lagged outcomes from 1982, 1987, and 1992, as well as the share of four-year-olds in school, the share of K-12 students in private school, the share of mothers and fathers with children less than five years old who have completed 12th grade, the share of mothers and fathers with children less than five years old who have completed at least four years of college, and current expenditure per pupil in public elementary and secondary schools.

and fathers with children less than five years old who have completed at least four years of college, and current expenditure per pupil in public elementary and secondary schools.[17]

As Table 4 shows, the ITT effects are mostly robust to alternative SCM specifications. The effect on high school graduation is statistically significant across all specifications, although including additional predictors (and only a few lagged outcomes) attenuates it some. The effect on associate degrees is always small but statistically significant (for the reason discussed in Section 5.2). The effects on bachelor's degrees are mostly statistically significant and qualitatively similar in magnitude. All of the employment effects are positive but statistically insignificant at conventional levels.

6 Comparing Georgia UPK to Related Programs

Georgia UPK's long-run impacts are broadly consistent with the impacts of other preschool programs in the literature. See Table 5 for a comparison of ATT effects on high school graduation between Georgia UPK, Boston UPK, Head Start, Perry Preschool, and Carolina Abecedarian. Perry Preschool stands out as having the largest impact at 16.5 percentage points. Recall that this intervention was highly intensive, baseline graduation rates were generally lower for 1960s pre-K cohorts, and counterfactual preschool options have improved over time. At 4.0 percentage points, Georgia UPK's impact is at neither the bottom nor the top of the distribution. It is a little smaller than the impact of Boston UPK, the most similar program in the table.

[17] The current expenditure data are drawn from the 1990 and 1996 editions of the NCES Digest of Education Statistics. I calculate the other predictors using data from the U.S. Census Bureau's Current Population Survey.

Table 5. Comparison of ATT Effects on High School Graduation Among Related Preschool Programs

Study	Program	Treated Pre-K Cohorts	Impact Estimate in Percentage Points (Standard Error)
This paper	Georgia UPK	1995-1998	4.0 (2.1)
Gray-Lobe et al. (2021)	Boston UPK	1997-2003	6.0 (3.0)
Bailey et al. (2021)	Head Start	1969-1984	2.4 (1.2)
Garces et al. (2002)	Head Start	1968–1981	3.7 (5.3)
Deming (2009)	Head Start	1984-1990	8.6 (3.1)
Belfield et al. (2006)	Perry Preschool	1962-1967	16.5 (8.4)
Campbell et al. (2012)	Abecedarian	1972-1982	6.8 (7.2)

Notes: Adapted from Gray-Lobe et al. (2021). In the third column, an individual's pre-K cohort is defined as their birth year plus four. Note though that for some programs in the table, children were treated before their "pre-K year." For comparability, I use the impact estimate from this paper that is scaled to account for program enrollment and out-migration from Georgia; as discussed in Section 5.1.3, I divide the DiD point estimate and standard error by .42. The Bailey et al. (2021) impact estimate is also a scaled ITT effect.

7 Conclusion

In 1995, Georgia became the first state to implement a statewide universal pre-K program. Since then, several others have followed suit, and UPK programs are frequently proposed at the federal, state, and local levels. However, few studies have evaluated their long-run impacts.

This paper provides new evidence on the long-run impacts of universal preschool using Georgia as a case study. I find evidence across two (albeit similar) identification strategies that Georgia UPK made children more likely to graduate from high school (by roughly 4.5 percent) and obtain a bachelor's degree (by roughly 13.7 percent). These are meaningful magnitudes, comparable to those from other well-known preschool programs. Some caution is warranted for both effects though. In particular, the effect on bachelor's degree attainment is statistically significant in the SCM analysis but not in the DiD analysis.

On the other hand, I find no effects on associate degree attainment and employment as an adult. The lack of an effect on associate degrees is tricky to interpret though; UPK might have induced some people to enroll in community college and others to forgo community college in favor of four-year college. The lack of evidence for an effect on employment is easier to interpret, although my analysis cannot rule out impacts on occupation, income, employment later in life, or other labor market outcomes. UPK's impacts on high school graduation and bachelor's degree attainment suggest that these other impacts may exist.

Notably, Georgia UPK produced impacts on educational attainment amidst a more modern preschool landscape and at a scale much larger than historical model programs. The program achieved these impacts despite overall program quality being only mediocre to good. These findings are informative to policymakers as they weigh the costs and benefits of new preschool programs. Large-scale programs might be expected to operate at quality levels similar to Georgia UPK. At the time of this writing, President Biden continues to push for a national UPK program that might operate very similarly to Georgia's. Though my results will not replicate exactly in other contexts, they can increase our confidence that large-scale universal preschool can improve the life trajectories of young children.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 2010, *105* (490), 493–505.

Bailey, Martha, Shuqiao Sun, and Brenden Timpe, "Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency," *American Economic Review*, 2021, *111* (12), 3963–4001.

Baker, Michael, Jonathan Gruber, and Kevin Milligan, "Universal Child Care, Maternal Labor Supply, and Family Well-Being," *Journal of Political Economy*, 2008, *116* (4), 709–745.

_, _, and _, "The Long-Run Impacts of a Universal Child Care Program," *American Economic Journal: Economic Policy*, 2019, *11* (3), 1–26.

Barnett, Steven, Kenneth Robin, Jason Hustedt, and Karen Schulman, "The State of Preschool 2003: State Preschool Yearbook," Technical Report, National Institute for Early Education Research, New Brunswick, NJ 2003.

Belfield, Clive, Milagros Nores, Steve Barnett, and Lawrence Schweinhart, "The High/Scope Perry Preschool Program: Cost-Benefit Analysis Using Data from the Age-40 Followup," *Journal of Human Resources*, 2006, *41* (1), 162–190.

Brackett, Margret, Gary Henry, and Jeanie Weathersby, "Report on the Expenditure of Lottery Funds Fiscal Year 1999," Technical Report, The Council for School Performance, GA 1999.

Bryan, Kris and Gary Henry, "Quality of Georgia's Pre-Kindergarten Program: 1997-98 School Year," Technical Report, Council for School Performance, Atlanta, GA October 1998.

Bryant, Donna, Margaret Burchinal, Lisa Lau, and Joseph Sparling, "Family and Classroom Correlates of Head Start Children's Developmental Outcomes," *Early Childhood Research Quarterly*, 1994, *9*, 289–309.

Campbell, Frances, Elizabeth Pungello, Margaret Burchinal, Kirsten Kainz, Yi Pan, Barbara Wasik, Oscar Barbarin, Joseph Sparling, and Craig Ramey, "Adult Outcomes as a Function of an Early Childhood Educational Program: an Abecedarian Project Follow-Up," *Developmental Psychology*, 2012, *48* (4), 1033–1043.

Cascio, Elizabeth, "Does Universal Preschool Hit the Target? Program Access and Preschool Impacts," *Journal of Human Resources*, 2021, pp. 1–60.

— **and Diane Schanzenbach**, "The Impacts of Expanding Access to High-Quality Preschool Education," *Brookings Papers on Economic Activity*, 2013, *44* (2), 127–178.

Conley, Timothy and Christopher Taber, "Inference with "Difference in Differences" with a Small Number of Policy Changes," *The Review of Economics and Statistics*, 2011, *93* (1), 113–125.

Deming, David, "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start," *American Economic Journal: Applied Economics*, 2009, *92* (4), 111–134.

Doudchenko, Nikolay and Guido Imbens, "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," Technical Report Working Paper 22791, National Bureau of Economic Research, Cambridge, MA October 2016.

Elango, Sneha, Jorge Luis García, James Heckman, and Andrés Hojman, "Early Childhood Education," in Robert Moffitt, ed., *Economics of Means-Tested Transfer Programs in the United States, Volume 2*, Chicago, IL: University of Chicago Press, 2016, chapter 4, pp. 235—-297.

Ferman, Bruno and Cristine Pinto, "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," *The Review of Economics and Statistics*, 2019, *101* (3), 452–467.

— **and** — , "Synthetic Controls with Imperfect Pretreatment Fit," *Quantitative Economics*, 2021, *12* (4), 1197–1221.

Fitzpatrick, Maria, "Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement," *The B.E. Journal of Economic Analysis Policy*, 2008, *8* (1), 1–40.

— , "Preschoolers Enrolled and Mothers at Work? The Effects of Universal Prekindergarten," *Journal of Labor Economics*, 2010, *28* (1), 51–85.

Friedman-Krauss, Allison, Steven Barnett, Karin Garver, Katherine Hodges, G.G. Weisenfeld, and Beth Gardiner, "The State of Preschool 2020: State Preschool Yearbook," Technical Report, National Institute for Early Education Research, New Brunswick, NJ 2021.

Garces, Eliana, Duncan Thomas, and Janet Currie, "Longer-Term Effects of Head Start," *American Economic Review*, 2002, *92* (4), 999–1012.

Gormley Jr., William and Ted Gayer, "Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program," *The Journal of Human Resources*, 2005, *40* (3), 533–558.

Gray-Lobe, Guthrie, Parag Pathak, and Christopher Walters, "The Long-Term Effects of Universal Preschool in Boston," Technical Report Working Paper 28756, National Bureau of Economic Research, Cambridge, MA May 2021.

Heckman, James, Seong Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz, "The Rate of Return to the High/Scope Perry Preschool Program," *Journal of Public Economics*, 2010, *94* (1-2), 114–128.

Henry, Gary, Bentley Ponder, Dana Rickman, Andrew Mashburn, Laura Henderson, and Craig Gordon, "An Evaluation of the Implementation of Georgia's Pre-K Program: Report of the Findings from the Georgia Early Childhood Study (2002-03)," Technical Report, Andrew Young School of Policy Studies, Georgia State University, Atlanta, GA December 2004.

_ , Craig Gordon, and Dana Rickman, "Early Education Policy Alternatives: Comparing Quality and Outcomes of Head Start and State Prekindergarten," *Educational Evaluation and Policy Analysis*, 2006, *28* (1), 77–99.

_ , _ , Laura Henderson, and Bentley Ponder, "Georgia Pre-K Longitudinal Study: Final Report, 1996-2001," Technical Report, Andrew Young School of Policy Studies, Georgia State University, Atlanta, GA May 2003.

_ , Laura Henderson, Bentley Ponder, Craig Gordon, Andrew Mashburn, and Dana Rickman, "Report on the Findings from the Early Childhood Study: 2001-02," Technical Report, Andrew Young School of Policy Studies, Georgia State University, Atlanta, GA August 2003.

Hill, Carolyn, William Gormley Jr., and Shirley Adelstein, "Do the short-term effects of a high-quality preschool program persist?," *Early Childhood Research Quarterly*, 2015, *32*, 60–79.

Johnson, Rucker and C. Kirabo Jackson, "Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending," *American Economic Journal: Economic Policy*, 2019, *11* (4), 310–349.

Kottelenberg, Michael and Steven Lehrer, "Targeted or Universal Coverage? Assessing Heterogeneity in the Effectsof Universal Child Care," *Journal of Labor Economics*, 2017, *35* (3), 609–653.

Murnane, Richard and John Willett, *Methods Matter: Improving Causal Inference in Educational and Social Science Research*, New York, NY: Oxford University Press, 2011.

Owen, Thompson, "Head Start's Long-Run Impact: Evidence from the Program's Introduction," *Journal of Human Resources*, 2018, *53* (4), 1100–1139.

Pages, Remy, Dylan Lukes, Drew Bailey, and Greg Duncan, "Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts," *Educational Evaluation and Policy Analysis*, 2020, *42* (4), 471–492.

Phillips, Deborah, Mark Lipsey, Kenneth Dodge, Ron Haskins, Daphna Bassok, Margaret Burchinal, Greg Duncan, Mark Dynarski, Katherine Magnuson, and Christina Weiland, "The Current State of Scientific Knowledge on Pre-Kindergarten Effects," Technical Report, Brookings Institution 2017.

Tietze, Wolfgang, Debby Cryer, Joachim Bairrão, Jesús Palacios, and Gottfried Wetzel, "Comparisons of Observed Process Quality in Early Child Care and Education Programs in Five Countries," *Early Childhood Research Quarterly*, 1996, *11*, 447–475.

van Huizen, Thomas and Janneke Plantenga, "Do children benefit from universal early childhood education and care? A meta-analysis of evidence from natural experiments," *Economics of Education Review*, 2018, *66*, 206–222.

Weiland, Christina and Hirokazu Yoshikawa, "Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills," *Child Development*, 2013, *84* (6), 2112–2130.

Woodyard, Henry, Tim Sass, and Ishtiaque Fazlul, "The Efficacy of School-Based Pre-K Program Sites in a Metro-Atlanta School District," Technical Report, Metro Atlanta Policy Lab for Education, Atlanta, GA July 2022.

Zerpa, Mariana, "Short and Medium Run Impacts of Preschool Education: Evidence from State Pre-K Programs," Technical Report, Mimeo February 2021.

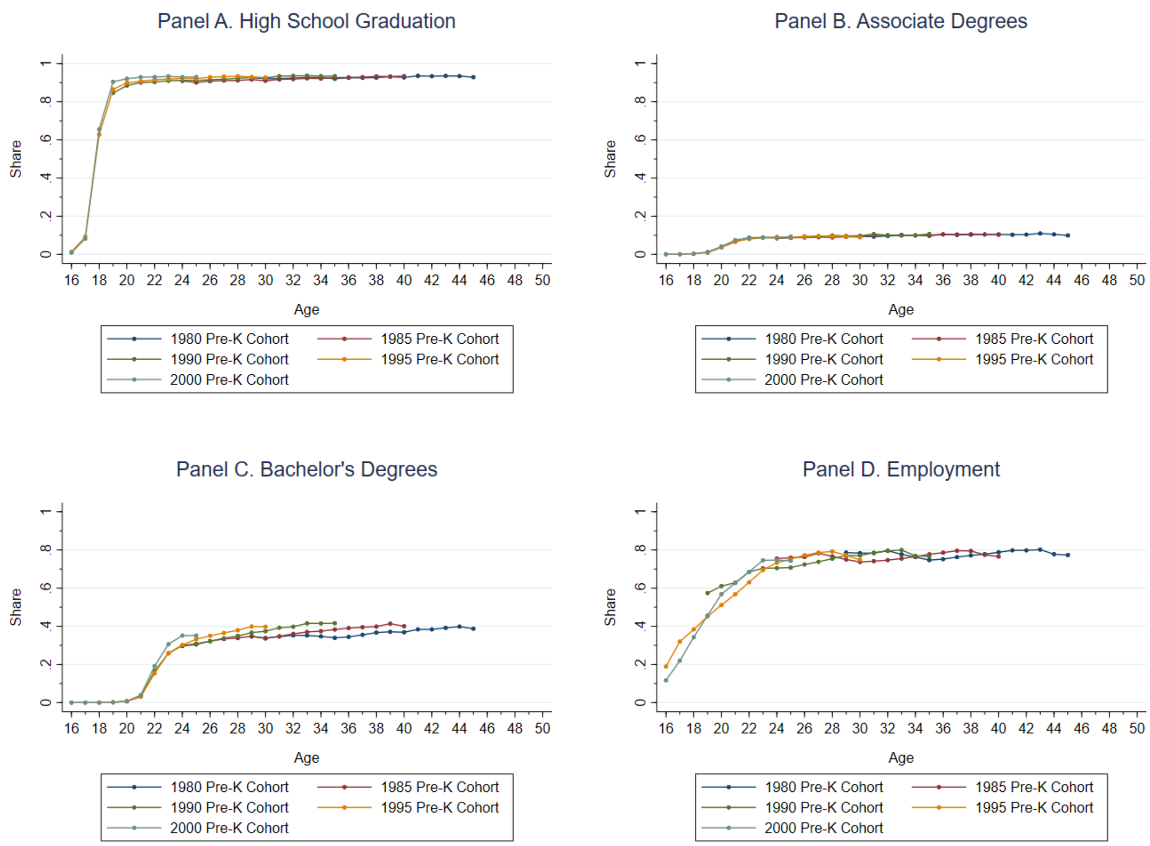
A Supplementary Tables and Figures

Table A1. Age Range and Sample Size of Pre-K Cohorts in the ACS Data

		Sample Size	
Pre-K Cohort	Observed Ages	Control Group States	Georgia
1980	(29, 45)	73,210	10,804
1981	(28, 44)	74,959	10,814
1982	(27, 43)	76,459	11,284
1983	(26, 42)	77,862	11,474
1984	(25, 41)	81,091	11,806
1985	(24, 40)	81,935	11,864
1986	(23, 39)	81,301	11,687
1987	(22, 38)	79,883	12,014
1988	(21, 37)	78,209	11,546
1989	(20, 36)	77,829	12,299
1990	(20, 35)	74,933	12,130
1991	(20, 34)	69,471	11,566
1992	(20, 33)	63,934	10,961
1993	(20, 32)	60,689	10,753
1994	(20, 31)	57,612	9,848
1995	(20, 30)	53,139	9,071
1996	(20, 29)	47,712	8,297
1997	(20, 28)	41,907	7,292
1998	(20, 27)	36,587	6,327

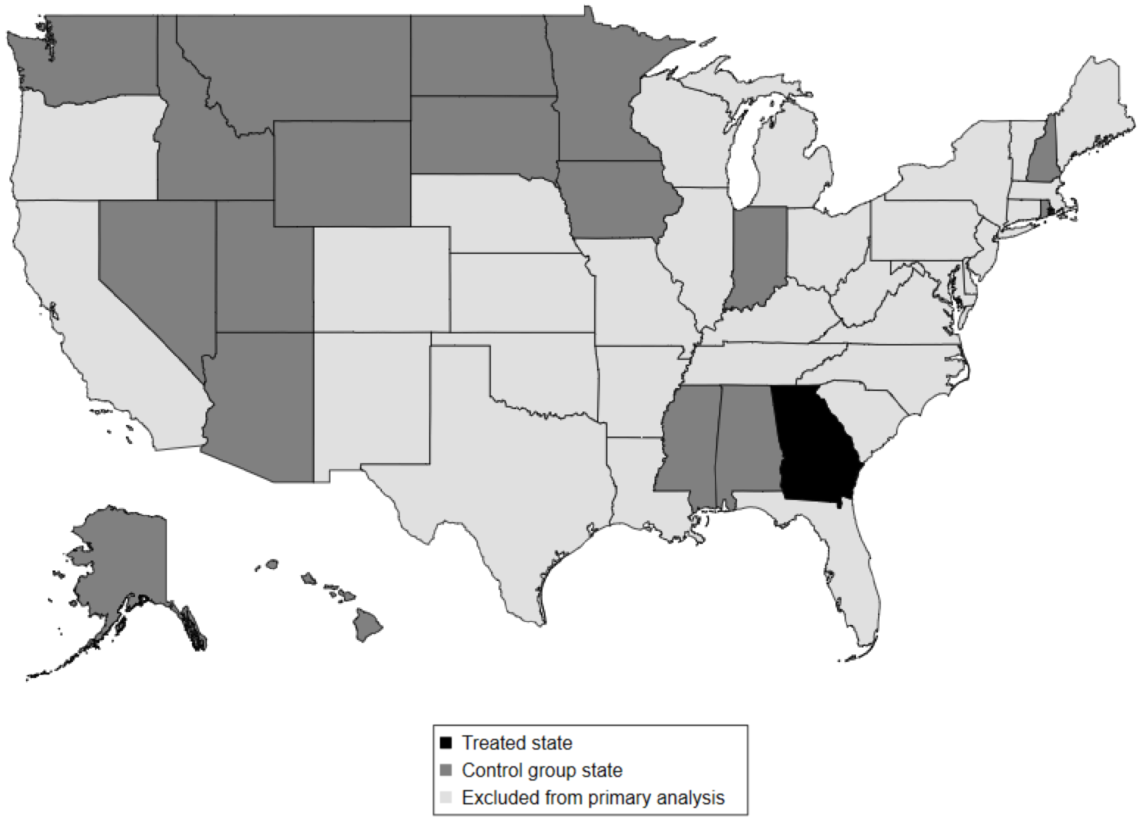
Notes: Author's calculations using American Community Survey data from 2005 through 2020. Following Zerpa (2021), the 18 control group states are Alabama, Alaska, Arizona, Hawaii, Idaho, Indiana, Iowa, Minnesota, Mississippi, Montana, Nevada, New Hampshire, North Dakota, Rhode Island, South Dakota, Utah, Washington, and Wyoming. Recall that the minimum age threshold for inclusion in the sample is greater than 20 for most outcomes.

Figure A1. Long-Run Outcome Stabilization, by Age



Notes: Author's calculations using American Community Survey data from 2005 through 2020.

Figure A2. Map of the Treatment and Control Group States



Notes: Following Zerpa (2021), the 18 control group states are Alabama, Alaska, Arizona, Hawaii, Idaho, Indiana, Iowa, Minnesota, Mississippi, Montana, Nevada, New Hampshire, North Dakota, Rhode Island, South Dakota, Utah, Washington, and Wyoming. These are the states that had no more than a very small statewide pre-K program during the time of my analysis.