The Effect of Early Childhood Programs on Third-Grade Test Scores: Evidence from Transitional Kindergarten in Michigan*

By JORDAN BERNE, BRIAN JACOB, TAREENA MUSADDIQ, ANNA SHAPIRO, AND CHRISTINA WEILAND†

Early childhood education (ECE) programs boost children's school readiness and can have long-run effects on participants' educational, income, and health outcomes (Phillips et al., 2017). Transitional Kindergarten (TK) programs are a relatively new entrant into the early education landscape, having come about as states pushed back the dates at which children can enter traditional kindergarten. Unlike many public early education options, TK programs are not targeted to children from families with low income. They also differ from many public ECE options in other important ways: TK operates only in public schools, is taught by state-certified teachers who are paid at the level of other K-12 educators, and tends to use a more academic-focused curriculum than many ECE programs (Shapiro et al., 2023).

The only TK program rigorously evaluated is the statewide program in California. Manship et al. (2017) found the program increased students' early literacy, language, and math skills along with kindergarten engagement. However, a study with a larger sample followed via administrative records found no benefits of TK on student's 3rd and 4th grade test scores (Lafortune and Hill, 2023).

We extend this literature by estimating the impacts of Michigan's TK program. In Michigan, TK funding is available for all children who turn five by the state's kindergarten birthday cutoff (September 1) or within three months after (September 2-December 1). Therefore, funding is available for all five-year-olds, but only the oldest four-year-olds. Hence, families of children who turn five *after* the state's September 1 kindergarten cutoff but *on or before* December 1 have three options: 1) enroll in TK (in districts that offer it and in which there is space); 2) waive into traditional K early (in all districts); or 3) choose an alternative public or private care setting, including staying home.

Leveraging the December 1 age-eligibility cutoff in a regression discontinuity (RD) framework, we find substantial positive impacts of attending TK on children's 3rd grade math and English Language Arts (ELA) test scores.

I. Data

This study uses longitudinal administrative data on the universe of Michigan public school students. Unfortunately, not every school district in Michigan has reliable data on TK enrollment. Therefore, we restrict our sample to districts that do not offer TK and districts with TK that reliably report student-level enrollment. For more information on how we identify districts with reliable data, and how these districts compare to those with unreliable data, see Appendix A.

We also limit our sample to students who turned five near December 1 in 2014 or 2018. (We exclude students in the intervening cohorts because their 3rd grade tests were disrupted by the COVID-19 pandemic.) Our other data cleaning decisions and sample restrictions are detailed in Appendix A.

* Go to https://doi.org/10.XXXX/pandp.2024XXXX to visit the article page for additional materials.

† Berne: University of Michigan (email: jsberne@umich.edu). Jacob: University of Michigan (email: bajacob@umich.edu). Musaddiq: Mathematica Inc. (email: tmusaddiq@mathematica-mpr.com). Shapiro: RAND Corporation (email: ashapiro@rand.org). Weiland: University of Michigan (email: weilandc@umich.edu). The data used for this research was structured by the MERI-Michigan Education Data Center (MEDC), and is not identical to data collected and maintained by the Michigan Department of Education (MDE) or Michigan's Center for Educational Performance and Information (CEPI). Results and opinions presented here do not reflect the views or positions of MDE or CEPI. This research was funded by the Smith Richardson Foundation, as well as training grants R305B20011 and R305B170015 from the U.S. Department of Education's Institute of Education Sciences. We thank MDE and CEPI for their partnership, especially Thomas Howell and Richard Lower. We also thank Katia Cordoba Garcia and Samuel Owusu for excellent research assistance and Jasmina Camo-Biogradlija for assistance with data agreements and project management.

TABLE 1—SUMMARY STATISTICS

	Non-TK Districts		TK Districts		
	All Students	Early K Students	All Students	Early K Students	TK Students
Female (%)	50	56	50	57	50
White (%)	48	33	74	63	77
Black (%)	36	51	12	18	11
Hispanic (%)	10	9	7	7	7
Economically disadvantaged (%)	70	76	46	56	38
Neighborhood median household income ($)	49,666	45,951	66,494	61,976	69,120
Charter school (%)	30	46	3	4	5
District is in city (%)	40	49	21	35	19
District is in suburb (%)	29	32	54	49	57
District is in rural area (%)	23	13	13	7	13
District mean 3rd grade math score (SD)	−0.264	−0.339	0.249	0.191	0.231
Observations	9,902	8,410	2,043	923	1,689

Note: "Early K" students are those who use a waiver to enroll in regular kindergarten despite turning five after the kindergarten birthday cutoff. We use the sample of students born within 30 days of the TK cutoff to construct these statistics. All statistics are calculated at the student level. See Appendix A for a summary statistics table with a larger set of characteristics.

As Table 1 shows, students in TK districts were more likely to be White, less likely to be economically disadvantaged, and more likely to live in suburbs. Compared to other students in the same district, TK students tended to be slightly more economically advantaged and more likely to live in a suburb. For more information on the districts with TK programs and the students who choose to participate, see Shapiro et al. (2023).

II. Empirical Strategy

Identification in this setting is challenging because at the cutoff (December 1), children not only become eligible for TK but also gain the ability to enter kindergarten early via a waiver (we denote students who choose to enter kindergarten early as EK). With the standard RD assumptions, we can only identify the combined effect of becoming eligible for *both* TK and EK.[1] Decomposing this intent-to-treat (ITT) effect into separate TK and EK effects requires more structure.

[1]To economize on space, we present results from the usual RD validity checks in Appendix C. A couple results raise concerns, but overall we believe we have a valid natural experiment.

As we show in Appendix B, the ITT effect for TK districts is a weighted combination of the TK and EK local average treatment effects (LATEs):

$$ITT = \Omega_{TK} LATE_{TK} + \Omega_K LATE_{EK} \quad (1)$$

where each Ω_x weight is the share of students at the cutoff who are compliers for option x. Notice that the quantities ITT, Ω_{TK}, and Ω_{EK} are all identified simply using districts that offer TK.

At this point, we have one equation with two unknowns: $LATE_{TK}$ and $LATE_{EK}$. Intuitively, since the ITT effect is a combination of two unobserved LATEs, recovering one of the two would allow us to back out the other. Our strategy is to use districts that do not offer TK as a second source of information to recover the LATE for the early K treatment.

In districts without TK, the RD cutoff generates variation in early K entry but not TK enrollment, allowing us to cleanly identify $LATE_{EK}$. However, using non-TK districts to infer $LATE_{EK}$ in TK districts requires some type of restriction on treatment effect heterogeneity. In theory, $LATE_{EK}$ may differ across districts. As we showed in Table 1, early kindergarten students in TK and non-TK districts differ on observable characteristics, and they likely differ in

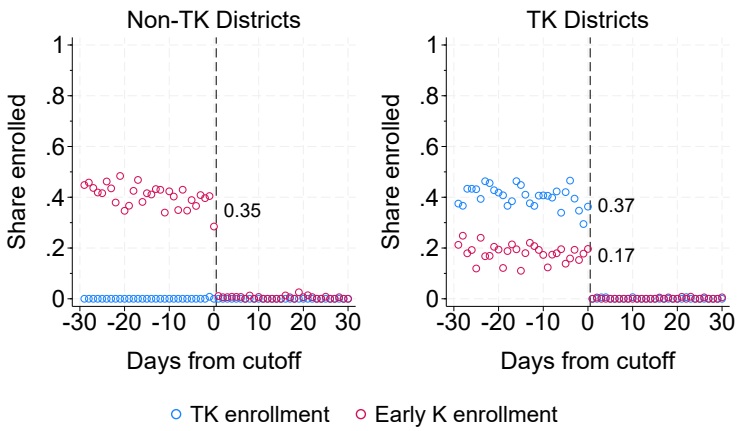


FIGURE 1. FIRST STAGE EFFECTS ON TK AND EK ENROLLMENT IN ONE'S PRE-K YEAR

unobservable ways too. We take two approaches to estimation in order to address the potential for treatment effect heterogeneity.

A. *Baseline Estimation Approach*

In our baseline approach, we assume the treatment effect of TK is the same in TK and non-TK districts, and estimate $LATE_{TK}$ and $LATE_{EK}$ jointly using both TK and non-TK districts. For student i in district d from cohort c, we estimate the following system of equations via two-stage least squares (2SLS):

$$Y_i = \beta_0 + \beta_1 TK_i + \beta_2 EK_i + f(dob_i) + \Pi \mathbf{X}_i + \lambda_{dc} + \varepsilon_{idc} \quad (2)$$

$$TK_i = \delta_0^{TK} + \delta_1^{TK} Left_i + \delta_2^{TK} Left_i \times DistHasTK_{dc} + f^{TK}(dob_i) + \Psi^{TK}\mathbf{X}_i + \theta_{dc}^{TK} + \varepsilon_{idc}^{TK} \quad (3)$$

$$EK_i = \delta_0^{EK} + \delta_1^{EK} Left_i + \delta_2^{EK} Left_i \times DistHasTK_{dc} + f^{EK}(dob_i) + \Psi^{EK}\mathbf{X}_i + \theta_{dc}^{EK} + \varepsilon_{idc}^{EK} \quad (4)$$

where $\mathbf{X}_i$ is a vector that includes student sex, race, and economic disadvantage status; λ_{dc}, θ_{dc}^{TK}, and θ_{dc}^{EK} are district×cohort fixed effects; and dob_i is date of birth. The f functions allow different linear relationships between date of birth and outcomes on either side of the cutoff and across districts with and without TK. The excluded instruments, $Left_i$ and $Left_i \times DistHasTK_{dc}$, indicate being born to the left of the RD cutoff and being left of the cutoff in a district×cohort in which TK is offered. We estimate these equations using a bandwidth of ±30 days. Our estimates for $LATE_{TK}$ and $LATE_{EK}$ are $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.[2]

B. *Preferred Estimation Approach*

In our preferred approach, we allow $LATE_{EK}$ to differ by student demographic characteristics. To operationalize this idea, we estimate early K effects in the *non-TK sample* separately for 8 demographic cells defined by sex × race (White or Asian vs. other races) × economic disadvantage status (disadvantaged vs. not). We estimate these effects using a 2SLS model analogous to Equations 2 and 4.

Next, within the *TK sample*, we calculate the share of early K students that belong to each of the same 8 demographic cells. We then use these shares as weights to aggregate the cell-specific effects from the non-TK sample. The result is a single $LATE_{EK}$ estimate that reflects the cell-specific treatment effects from the non-TK districts weighted to reflect the demographic composition of early K students in TK districts.

Once we have a "demographically-adjusted" estimate of $LATE_{EK}$, it is straightforward to

[2] As a check on the 2SLS specification, we also estimate the components of Equation 1 separately and back out $LATE_{TK}$. The estimates are nearly identical.

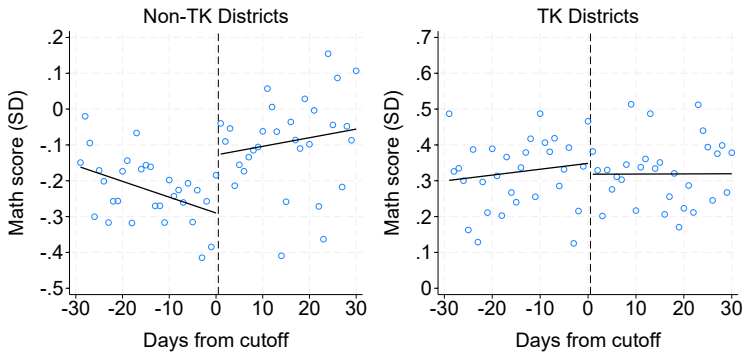


FIGURE 2. INTENT-TO-TREAT EFFECTS ON 3RD GRADE MATH SCORES

back out $LATE_{TK}$ using Equation 1 and estimates of ITT, Ω_{TK}, and Ω_{EK} from the TK sample.

We conduct inference via bootstrap because this approach contains multiple steps. For consistency, we bootstrap in the baseline approach too. See Appendix E for more details.[3]

III. Results

Figure 1 shows we have strong first stage impacts on TK and early K enrollment at the RD cutoff. In non-TK districts, 35 percent of students waive into kindergarten early. In TK districts, 37 percent of students enroll in TK and 17 percent enter kindergarten early.

Figure 2 presents a visual representation of our ITT effects on 3rd grade math scores, which also provides intuition for our estimation of the treatment effects of TK and early K. In non-TK districts, there is a large negative discontinuity at the cutoff. Because early K is the only option associated with the cutoff in these districts, the figure tells us that students who waive into K early score substantially lower on the 3rd grade math exam. This result is not surprising given that early K students take 3rd grade tests one year earlier than they otherwise would, giving them one less year of cognitive development (Deming and Dynarski, 2008). On the other hand, in TK districts—where the cutoff is associated with both K waiving and TK—there is hardly any discontinuity at the cutoff. If early K has a negative effect in TK districts like it does in non-TK districts, this implies TK must have an offsetting positive effect.

Consistent with this visual intuition, the results in Table 2 indicate that TK enrollment leads to substantial improvements in 3rd grade math scores. Our baseline estimate is 0.21 standard deviations and our preferred estimate is 0.29 standard deviations. The reason for the difference is that the demographic subgroups that dominate the early K compliers in TK districts (e.g., White and Asian girls regardless of economic disadvantage and White and Asian boys who are not economically disadvantaged) experience the largest negative effects from attending kindergarten early (see Appendix D). Using our preferred approach, we estimate that TK enrollment increases 3rd grade ELA scores by 0.19 standard deviations, although this estimate is not statistically distinguishable from 0.

Several points are worth noting when interpreting these estimates. First, the TK impacts are relative to a counterfactual of starting kindergarten "on schedule" or later, having spent one's pre-K year receiving care at home, attending Michigan's income-targeted public pre-K program, attending private pre-K, or spending the year in some other type of arrangement. Second, and related, because they reflect the impact for students close to the December 1 cutoff, our estimates capture the effects of TK for children who are among the oldest in their birth cohort.

In Appendices C and D, we conduct RD validity checks and elaborate on our identification strategies.

[3]Table 2 summarizes our inference results with p-values rather than standard errors because the bootstrap distributions in our preferred approach are non-normal and contain extreme outliers, rendering the standard errors uninformative.

TABLE 2—IMPACTS OF TK AND EK ON 3RD GRADE TEST SCORES

	Math				ELA			
	Baseline		Preferred		Baseline		Preferred	
$LATE_{TK}$	0.252**	0.212*	0.331*	0.294	0.123	0.097	0.209	0.191
[P-value]	[0.046]	[0.051]	[0.088]	[0.111]	[0.321]	[0.401]	[0.253]	[0.293]
$LATE_{EK}$	−0.378***	−0.366***	−0.557*	−0.557*	−0.240*	−0.219*	−0.435	−0.435
[P-value]	[0.000]	[0.000]	[0.092]	[0.092]	[0.061]	[0.078]	[0.181]	[0.181]
Controls		X		X		X		X
Control mean		0.302				0.286		
Observations		15,680				15,669		

Note: Inference is conducted via bootstrap, with clustering on the running variable. Appendix E elaborates on our inference strategy. In the preferred approach, we always exclude controls when estimating EK LATEs because the demographic subgroups are defined by the covariates. Hence, the EK estimates in the with- and without-controls columns are identical by construction. The control mean is the average of the outcome variable for students in TK districts born 1-5 days after the cutoff. The first stage F-statistics in the baseline model are 336.21 (math) and 339.82 (ELA) for TK and 400.34 (math) and 340.52 (ELA) for EK. *** p<0.01, ** p<0.05, * p<0.1

IV. Discussion

Research commonly finds that attending preschool has sizable impacts on kindergarten readiness, followed by partial or complete fade out in the early elementary years, and then re-emergence of benefits in early adulthood (Phillips et al., 2017). Our finding that Michigan TK raises 3rd grade math scores stands in stark contrast to this typical pattern.

The magnitude of our estimates is large relative to the prior literature. Across all relatively rigorous evaluations of programs since the 1960s, the average impact of preschool on children's *end-of-preschool* cognitive skills is about 0.25 standard deviations (Duncan and Magnuson, 2013). Our impact estimates are the same size for students in 3rd grade. To give another reference point, our math estimates amount to 57% of expected cognitive development between 3rd and 4th grade (Hill et al., 2008).

The more positive impacts of Michigan TK compared with other ECE programs could be due to a variety of factors, including (a) less formal early childhood education among the control group in our sample, (b) the quality of TK teachers, and (c) better alignment between TK and early elementary schooling. In future work, we will investigate these and other potential mechanisms as well as explore the heterogeneity of treatment effects across subgroups. We hope that our findings will inform the development and scaling of other successful ECE models in Michigan and elsewhere.

REFERENCES

Deming, David, and Susan Dynarski. (2008). "The Lengthening of Childhood." *Journal of Economic Perspectives* 22 (3): 71–92.

Duncan, Greg, and Katherine Magnuson. (2013). "Investing in Preschool Programs." *Journal of Economic Perspectives* 27 (2): 109-132.

Hill, Carolyn, Howard Bloom, Alison Black, and Mark Lipsey. (2008). "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2 (3): 172-177.

Lafortune, Julien, and Laura Hill. 2023. "Assessing Transitional Kindergarten's Impact on Elementary School Trajectories." Public Policy Institute of California.

Manship, Karen, Heather Quick, Aleksandra Holod, Nicholas Mills, Burhan Ogut, Jodi Jacobson Chernoff, Jarah Blum, Alison Hauser, Jennifer Anthony, and Raquel González. (2015). "Impact of California's Transitional Kindergarten Program, 2013–14." American Institutes for Research.

Phillips, Deborah, Mark Lipsey, Kenneth Dodge, Ron Haskins, Daphna Bassok, Margaret Burchinal, Greg Duncan, Mark Dynarski, Katherine Magnuson, and Christina Weiland. (2017). "Puzzling it out: The Current State of Scientific Knowledge on Pre-

Kindergarten Effects - A Consensus Statement." The Brookings Institution.

Shapiro, Anna, Jordan Berne, Katia Cordoba Garcia, Brian Jacob, Tareena Musaddiq, Samuel Owusu, and Christina Weiland. (2023). "Michigan Transitional Kindergarten: A First Look at Program Reach and Features." University of Michiga, Education Policy Initiative.

Online Appendix for The Effect of Early Childhood Programs on Third-Grade Test Scores: Evidence from Transitional Kindergarten in Michigan

Jordan Berne[1], Brian Jacob[1, 2], Tareena Musaddiq[3], Anna Shapiro[4], and Christina Weiland[1]

[1]University of Michigan
[2]National Bureau of Economic Research
[3]Mathematica Inc.
[4]RAND Corporation

January 2024

Contents

A Sample and Variable Construction **3**
A.A Identifying TK and Non-TK District×Cohorts . . . 3
A.B Student-Level Sample Restrictions . . . 4
A.C Constructing Grade and District Variables . . . 4
A.D Summary Statistics . . . 5

B Derivation of Equation 1 **7**

C RD Validity Checks **8**
C.A Attrition . . . 8
C.B Density Manipulation . . . 8
C.C Covariate Continuity . . . 10

D Robustness **13**
D.A Models With and Without Covariates . . . 13
D.B Subgroup Estimates in the "Preferred Approach" . . . 13
D.C Bounding the TK LATE . . . 14
D.D Alternative Identification Assumptions . . . 15

E Inference **16**

A Sample and Variable Construction

A.A Identifying TK and Non-TK District×Cohorts

Identifying which districts and charter schools offered TK in a particular year is not always straightforward. In the administrative data, all kindergarteners and TK students are marked as being in grade 0. A separate flag is meant to indicate TK enrollment. However, because TK students are funded as regular kindergarteners, districts were advised but not required to use the separate TK flag during the time of our study. As a result, in many districts we cannot tell whether a student in grade 0 is in TK or traditional kindergarten.

To minimize measurement error in program enrollment, we limit our sample to districts and cohorts in which we are confident about individual-level TK data. When a district reports at least 10 TK students in the data in a given year, we are reasonably confident the district offered TK that year. For one, we believe districts are unlikely to mistakenly report 10 or more students as being in TK. Second, as we discuss further below, grade progression patterns match our expectations in districts that meet our data reliability standards, but not in districts that don't. This restriction may drop districts with particularly small TK programs, but it increases the likelihood that included districts are categorized accurately.

Throughout the paper, when we refer to TK districts, we are referring to district×years that report 10+ TK students. Because a given district may report 10+ TK students in one year but not another, TK districts are identified at the district×year level. After identifying TK districts, we then define the sample of students in TK districts. This sample consists of students in districts that offered TK the year before one's "scheduled" kindergarten year (based on students' birthdays and the statewide kindergarten cutoff). Thus, the sample of students in TK districts is defined at the district×cohort level.

We are also reasonably confident that a certain set of districts never offered TK. Throughout the paper, when we refer to non-TK districts, we are referring to districts that meet two conditions. First, a district must not have reported a single TK student in the data in any year. Second, a district must not have had TK in school year 2021-22 based on an extensive data triangulation process our team conducted that year. In spring 2022, our team reviewed district websites and communicated with district staff via email and phone calls to make a determination for every district about whether they offered TK in that school year.

After identifying TK and non-TK districts, we impose an additional district-level sample restriction that comes from our broader project evaluating Michigan TK. In the broader project, we are interested in estimating treatment effect heterogeneity across districts. For that purpose, we focus on districts with a positive, precise discontinuity in TK enrollment at the RD cutoff—either on their own or when pooled with observably similar districts. This restriction eliminates very small districts and larger districts with no discernible discontinuity at the cutoff. Overall, the restriction only drops 3.2% of students from the TK district sample. We impose this restriction in this paper too for sample consistency across our projects.

Ultimately, our analysis sample contains 292 district×cohorts from 205 TK districts and 696 district×cohorts from 376 non-TK districts. Note that the number of non-TK *district×cohorts* is not exactly twice the number of non-TK *districts* because some small districts do not have students born within the 30-day bandwidth in both cohorts.

Examining the grade progression patterns of TK and early K students provides reassurance that our TK and non-TK districts are categorized correctly and that student-level program enrollment is accurate. 77% of early kindergarten students in non-TK districts and 82% of early kindergarten students in TK districts move on to 1st grade the year after waiving into kindergarten. On the other

hand, in districts that don't meet our requirements for reliable reporting, only 51% of students who appear to have waived into kindergarten early move on to 1st grade the following year. We believe it likely that many of these students were actually enrolled in TK, which would explain why a relatively low share advances to 1st grade the next year. In our TK sample, 98% of students enrolled in TK move on to traditional kindergarten the following year.

A.B Student-Level Sample Restrictions

Because our paper focuses on the effects of early learning programs, we drop students who attended neither TK nor kindergarten in a Michigan public school. This restriction drops Michigan students observed in later grades who attended early grades in private schools, home schools, or schools outside Michigan. This is a relatively small group; for example, such students constituted roughly 8% of Michigan third-graders in school year 2018-19.

Relatedly, we also drop students who attended neither TK nor kindergarten in a Michigan public school for at least 20 days. This sample restriction is meant to ensure that "treated" students experienced at least some treatment. This 20-day threshold drops 0.6% of all students (relative to a 0-day threshold).

Lastly, to accommodate our regression discontinuity research design, we drop students with invalid birthday information. This includes students with no birthday information, multiple listed birthdays, and birthdays that seem implausible.[1] All together, we drop less than 0.1% of students due to birthday-related reasons.

A.C Constructing Grade and District Variables

Within the set of TK and non-TK districts with reliable TK information, it is straightforward to distinguish between TK, EK, and "on-schedule" kindergarten students. Students in all three of these groups are marked as being in grade 0. TK students are identified using a separate flag for enrollment in a TK program. Among the remaining "grade 0" students, we can distinguish between early and on-schedule K students using the observed year of K enrollment, students' birthdays, and institutional knowledge of Michigan's kindergarten cutoff dates.

Some students appear in the data multiple times in the same school year, usually because they enrolled in different grades or schools in the same year. We clean the data so that each student is assigned to a single grade×school in a given school year. When a student is observed in the same grade multiple times in a year, we keep the observation with the most days attended. When a student is observed in multiple grades K or above in the same year, we again keep the observation with the most days attended. However, when students are observed in more than one of TK, K, or an early childhood program in the same year, we use a more nuanced procedure that uses days attended and grade progression to assign them to a single grade/program.

Among all TK students in our analysis sample, 3% are also enrolled in K in the same year. For these students, we use information on days attended and grade progression to determine whether to keep their TK or K observation. Specifically, we use the following algorithm:

- For students who attend TK first, followed by K, in year t:

[1] We consider a birthday implausible if it implies a student was born outside the two-year window that would be expected based on the year a student first enrolls in TK or K. The exact length of the window varies slightly by cohort to accommodate Michigan's changing kindergarten entry policies, but each window accounts for kindergarten redshirting and early entry. We also add a one-month cushion to the front and back ends of each window to account for non-compliance.

  - If they attend TK in year $t+1$, we keep the TK observation in year t
  - If they attend K in year $t+1$, we keep the observation in year t with more days attended. Note: if a year t observation has a missing value for days attended, we do not keep that observation.
  - If they attend 1st or 2nd grade in year $t+1$, we keep the K observation in year t

- For students who attend K first, followed by TK, in year t:
  - If they attend TK or K in year $t+1$, we keep the TK observation in year t
  - If they attend 1st or 2nd grade in year $t+1$, we keep the K observation in year t

After reconciling students enrolled in TK and K in the same year, we then reconcile students enrolled in TK or K in the same year as an early childhood program. 0.4% of all TK students and 0.4% of all K students are also enrolled in an early childhood program in the same year. We cannot use days *attended* for this reconciliation because it is not available in our data for early childhood programs. Instead, we compare the number of days students were *enrolled* for, keeping the observation with the higher number. For TK and K, we have data on the number of school days enrolled. For early childhood programs, the enrollment variable includes weekends, so we multiply by 5/7 to make it comparable with the TK and K variable. When a student is enrolled in multiple early childhood programs in the same year, we use the observation with the most days enrolled.

Once the data is unique at the student×year level, we assign each student to the district they are first observed in. The idea is to capture the district a student could have enrolled in when they were on the margin of age-eligibility for TK and early K entry. Students who participate in TK are assigned to their TK district, and other students are assigned to their kindergarten district.

A.D Summary Statistics

Table A1 presents summary statistics for a broader range of student-, school-, and district-level characteristics than Table 1 in the paper. The overall takeaways are much the same.

Table A1: Summary Statistics

	All Students		TK Students	Early K Students	
	TK Districts	Non-TK Districts	TK Districts	TK Districts	Non-TK Districts
Female (%)	50	50	50	57	56
White (%)	74	48	77	63	33
Black (%)	12	36	11	18	51
Hispanic (%)	7	10	7	7	9
Asian American (%)	6	4	5	11	6
Other race (%)	1	2	1	1	1
Economically disadvantaged (%)	46	70	38	56	76
Prior state pre-K enrollment (%)	3	4	7	13	13
LEP status (%)	10	11	6	21	15
Neighborhood White share (%)	85	64	86	81	51
Neighborhood poverty share (%)	9	18	8	13	21
Neighborhood unemployment rate (%)	20	13	20	19	13
Neighborhood BA attainment rate (%)	20	13	20	19	13
Neighborhood median household income ($)	66,494	49,666	69,120	61,976	45,951
School is in a city (%)	20	38	19	33	47
School is in a suburb (%)	50	31	50	46	34
School is in a town (%)	12	7	12	8	5
School is in a rural area (%)	18	23	20	13	14
Magnet school (%)	8	18	7	7	20
School enrollment (%)	444	443	433	441	479
School pupil:teacher ratio (%)	17	18	16	17	18
School FRL share (%)	43	67	41	49	73
Charter school (%)	3	30	5	4	46
District is in a city (%)	21	40	19	35	49
District is in a suburb (%)	54	29	57	49	32
District is in a town (%)	12	8	11	9	5
District is in a rural area (%)	13	23	13	7	13
District free- and reduced-price lunch share (%)	40	64	38	43	70
District LEP share (%)	7	11	6	12	11
District average 3rd grade math M-STEP score (SD)	0.249	−0.264	0.231	0.191	−0.339
Observations	9,902	8,410	2,043	923	1,689

Note: We use the sample of students born within 30 days of the TK cutoff to construct these statistics. All statistics are calculated at the student level.

B Derivation of Equation 1

In this section, we derive the expression for the intent-to-treat (ITT) effect shown in Equation 1 in the paper. The equation shows that the ITT effect is a weighted average of the TK and EK local average treatment effects (LATEs). For the sake of readability in this appendix, we'll use slightly different notation than in the paper:

- L_i is an indicator for being born to the left of December 1st (i.e., on or before).
- Treatment status, D_i, may take on values TK for TK, EK for waiving into K early, and 0 for doing neither TK nor waiving into K.
- $D_i(1)$ is the treatment a student would choose if they're to the left of the cutoff; $D_i(0)$ is what they would choose if they're to the right.
- Ω_x is the share of students who would participate in treatment x when eligible for all treatments, where x takes on values TK, EK, and 0 (neither TK nor EK).
- Y_i is a student's observed outcome and $Y_i(D)$ is their potential outcome under treatment D.

Now let's derive Equation 1. Focusing only on district×cohorts with TK, the ITT effect of being to the left of the cutoff can be written as:

$$ITT = E[Y_i|L_i = 1] - E[Y_i|L_i = 0]$$

We can break this equation apart by program complier types:

$$ITT = \underbrace{\Omega_{TK}E[Y_i|D_i(1) = TK, L_i = 1] + \Omega_{EK}E[Y_i|D_i(1) = EK, L_i = 1] + \Omega_0 E[Y_i|D_i(1) = 0, L_i = 1]}_{\text{Left of cutoff}}$$
$$- \underbrace{\Omega_{TK}E[Y_i|D_i(1) = TK, L_i = 0] - \Omega_{EK}E[Y_i|D_i(1) = EK, L_i = 0] - \Omega_0 E[Y_i|D_i(1) = 0, L_i = 0]}_{\text{Right of cutoff}}$$

The IV exclusion restriction implies $E[Y_i|D_i(1) = 0, L_i = 1] = E[Y_i|D_i(1) = 0, L_i = 0]$ because outcomes depend on treatment, not treatment eligibility. So, these terms cancel out and we have:

$$\begin{aligned} ITT = {} & \Omega_{TK}E[Y_i|D_i(1) = TK, L_i = 1] + \Omega_{EK}E[Y_i|D_i(1) = EK, L_i = 1] \\ & - \Omega_{TK}E[Y_i|D_i(1) = TK, L_i = 0] - \Omega_{EK}E[Y_i|D_i(1) = EK, L_i = 0] \end{aligned}$$

Rearranging and substituting in potential outcomes, we have:

$$\begin{aligned} ITT = {} & \Omega_{TK}\{E[Y_i|D_i(1) = TK, L_i = 1] - E[Y_i|D_i(1) = TK, L_i = 0]\} \\ & + \Omega_{EK}\{E[Y_i|D_i(1) = EK, L_i = 1] - E[Y_i|D_i(1) = EK, L_i = 0]\} \end{aligned}$$

$$\begin{aligned} ITT = {} & \Omega_{TK}\{E[Y_i(TK)|D_i(1) = TK] - E[Y_i(0)|D_i(1) = TK]\} \\ & + \Omega_{EK}\{E[Y_i(EK)|D_i(1) = EK] - E[Y_i(0)|D_i(1) = EK]\} \end{aligned}$$

$$ITT = \Omega_{TK}\underbrace{E[Y_i(TK) - Y_i(0)|D_i(1) = TK]}_{LATE_{TK}} + \Omega_{EK}\underbrace{E[Y_i(EK) - Y_i(0)|D_i(1) = EK]}_{LATE_{EK}}$$

C RD Validity Checks

C.A Attrition

Intuitively, the RD analysis requires that students born just before and just after December 1 be similar in ways other than treatment eligibility. Sample attrition poses a potential threat to this fundamental assumption. If the type of students who leave the sample are systematically different on either side of the cutoff, the resulting sample may not be continuous in observable or unobservable characteristics through the cutoff.

Attrition may occur for two reasons in our context. First, students may exit our data because they leave the Michigan public school system before 3rd grade. Second, students enrolled in a Michigan public school in 3rd grade may not have test score information in the data.

Table A2 shows that there does not appear to be differential attrition at the cutoff. In both TK and non-TK districts, around 93% of students in our sample born near December 1 are also observed in 3rd grade in a later year. The difference in this likelihood at the cutoff is small and statistically insignificant. When we account for missing test score data, the probability of remaining in a Michigan public school and having 3rd grade test score data is around 87% for TK and non-TK districts. Again, there is no evidence that this attrition occurs differentially at the cutoff.

Table A2: Attrition Estimates

	Control Mean	Estimate	Standard Error	P-value
Panel A. Non-TK Districts				
Ever observed in 1st grade	0.980	−0.009	0.004	0.022
Ever observed in 2nd grade	0.956	−0.009	0.006	0.130
Ever observed in 3rd grade	0.931	−0.005	0.008	0.544
Number of grades observed in between 1st and 3rd	2.87	−0.020	0.010	0.133
Has a 3rd grade math test score	0.856	−0.002	0.015	0.912
Has a 3rd grade ELA test score	0.857	−0.003	0.016	0.830
Panel B. TK Districts				
Ever observed in 1st grade	0.982	−0.008	0.004	0.069
Ever observed in 2nd grade	0.956	−0.004	0.005	0.418
Ever observed in 3rd grade	0.927	−0.003	0.006	0.598
Number of grades observed in between 1st and 3rd	2.86	−0.020	0.010	0.240
Has a 3rd grade math test score	0.877	−0.015	0.017	0.394
Has a 3rd grade ELA test score	0.876	−0.014	0.017	0.398

Note: A student is coded as not having a 3rd grade test score if they are not observed in 3rd grade or if they are observed in 3rd grade but do not have a test score in their first year of 3rd grade.

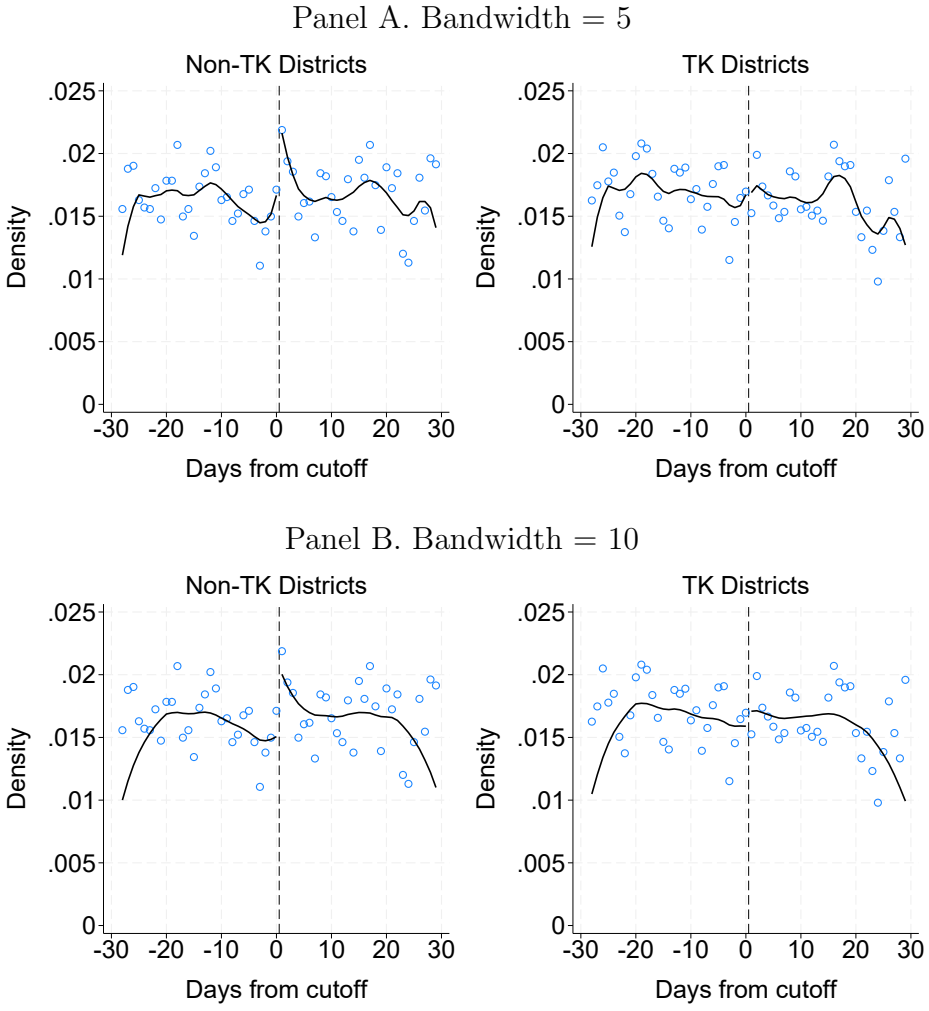
C.B Density Manipulation

As another check of the RD assumptions necessary for causal inference, we investigate whether the density of our sample is continuous through the cutoff. In our context, it seems unlikely that families manipulate the running variable (i.e., children's birthdays) in order to gain access to TK or EK—either through misreporting or birth timing. The more plausible concern is that families with children born between September 2 and December 1 may relocate to districts that offer TK

to gain access to the program. If this were the case, the children to the left of the cutoff may be systematically different than those to the right within district type.

We check for potential discontinuities in sample density using two approaches. First, we use the McCrary (2008) test with bandwidths of 5 and 10 days from the cutoff. Second, we use the Cattaneo et al. (2020) test that uses a mean squared error minimizing selection procedure to determine an optimal bandwidth. Our results are shown visually in Figure A1 and summarized in Table A3.

Figure A1: Density by Birthday



Notes: The smoothed lines in each panel are estimated using the McCrary (2008) density test.

In TK districts, the McCrary tests do not find a statistically significant discontinuity in density. The estimate from the Cattaneo, Jansson, and Ma test is statistically significant at the 10% level, but the magnitude of the discontinuity is quite small. On the other hand, our tests do find a potential discontinuity in non-TK districts. All of our tests find that the density is lower in non-TK districts to the left of the cutoff, i.e., for children who are age-eligible for TK and EK.

What might explain the density discontinuity in non-TK districts? As mentioned before, it's possible families with children born between September 2 and December 1 move from non-TK districts to TK districts to gain access to TK. Indeed, this would result in the density being lower to the left of the cutoff. However, if this were the story, the density would likely be lower for the entirety of the left-of-cutoff sample, whereas Figure A1 shows that the density is lower only for those near the

Table A3: Tests for Density Manipulation

		Non-TK Districts		TK Districts	
Test	Bandwidth	T-statistic	P-value	T-statistic	P-value
McCrary (2008)	5	−2.37	0.012	0.44	0.622
McCrary (2008)	10	−3.78	0.000	−1.03	0.288
Cattaneo, Jansson, and Ma (2020)	22.5	−4.09	0.000	−1.82	0.069

Note: The bandwidths for the McCrary (2008) tests are user-specified. The bandwidth for the Cattaneo, Jansson, and Ma (2020) tests is determined via a mean squared error minimizing selection procedure. The procedure selects 22.5 for non-TK and TK districts.

cutoff. Moreover, if this were the story, we would expect a corresponding *increase* in density to the left of the cutoff in districts with TK. Figure A1 shows that this is not the case. More generally, it is difficult explain why the density appears to dip close to the cutoff but not further away. Perhaps the birthday cutoff is most salient for families with children closest to the cutoff, but on the other hand, it is the older children who are more likely to be prepared for TK or EK.

A discontinuity in density is only problematic for our analysis insofar as it reflects differences between students to the left and right of the cutoff that are systematically related to test scores. In the next section, we explore whether observable student characteristics are continuous through the RD cutoff, which is a simple test of whether the density discontinuity is non-random. In short, we find no evidence that the density discontinuity reflects systematic differences between students.

The density discontinuity in non-TK districts warrants some caution. However, given the lack of a coherent theoretical explanation for the non-TK discontinuity, the continuity of observable characteristics through the cutoff, the potential for the non-TK discontinuity to be driven by noise, and the lower level of concern about density discontinuity in TK districts, we view this issue as unlikely to bias our estimates.

C.C Covariate Continuity

One of the key assumptions underlying our RD analysis is that student characteristics unrelated to treatment are continuous, on average, through the RD cutoff. We evaluate this assumption by estimating RD models with student characteristics as the outcome variable. Consistent with our main impact models, we use a bandwidth of ±30 days and specify linear relationships between the running variable and outcome variable that may vary on either side of the cutoff. Our results are presented in Table A4.

Overall, Table A4 provides strong evidence that baseline student characteristics are balanced at the cutoff. The vast majority of estimates are small and indistinguishable from 0. Only a small number of estimates are statistically significant at conventional levels, as would be expected to happen by chance when testing so many hypotheses.

To facilitate a summary test of covariate continuity through the cutoff, we combine the observable student characteristics into a single summary statistic. Specifically, we construct a measure of predicted 3rd grade math scores and then estimate RD models using the predicted score as the outcome variable. To obtain the relationship between student characteristics and test scores, we estimate a linear regression with 3rd grade math scores as the outcome variable and the following variables as predictors: student demographics (sex, race, and economic disadvantage); neighborhood characteristics (White share, poverty share, BA attainment rate, and log of median household income); district characteristics (share of students eligibile for free- or reduced-price lunch, log of

student enrollment, and urbanicity level); and region of Michigan. In both TK and non-TK districts, the estimated discontinuity in predicted test scores is small and indistinguishable from 0 (see the first row of each panel in Table A4).

The only characteristic with a statistically significant discontinuity in TK and non-TK districts is "prior state pre-K enrollment." By this we mean enrollment in Michigan's income-targeted state pre-K program the year before a student is on the margin of being age-eligible for TK or early kindergarten. This discontinuity is expected given our knowledge of the Michigan pre-K landscape. Michigan's pre-K program, called the Great Start Readiness Program (GSRP), is intended for students who turn 4 years old by September 1. However, children who turn 4 between September 2 and December 1 are sometimes eligible to enroll in GSRP as 3-year-olds when space is available after initial enrollment. Therefore, GSRP enrollment has the same birthday cutoff as TK and EK, although it applies two years before "on-schedule" kindergarten enrollment instead of one year before.

Consistent with our understanding of the GSRP age-eligibility rules, Table A4 shows that students born after December 1 do not enroll in GSRP before their pre-K year. On the other hand, 8.8% of students at the cutoff in non-TK districts and 5.7% of students at the cutoff in TK districts enroll in GSRP before their pre-K year.

If GSRP enrollment as a 3-year-old has a non-zero impact on student outcomes in 3rd grade, the discontinuity in enrollment at the RD cutoff could bias our estimates. In our main impact models, our omission of GSRP as a treatment option implicitly assumes that enrolling in GSRP before one's pre-K year does not have an effect on test scores that persists to 3rd grade. We view this assumption as a reasonable benchmark. For one, the discontinuity in 3-year-old GSRP enrollment is not particularly large, meaning GSRP's impact would have to be especially large to affect our estimates. Second, compared to TK and EK, the curriculum used in GSRP is typically less focused on academics and its teacher workforce is paid substantially less, making it plausible that test score impacts do not persist through 3rd grade. Third, the impacts of GSRP as a 3-year-old would have to persist conditional on the various child care and preschool arrangements children experience the following year, i.e., in their pre-K year. Assuming that potential test score impacts do not persist through 3rd grade is consistent with RCT evidence from the federal Head Start Impact Study, which found that cognitive impacts from 3-year-old Head Start enrollment did not persist through kindergarten (Puma et al., 2012).

Returning to our discussion on density manipulation from the previous section, this analysis of covariate continuity provides strong reassurance that the density discontinuity in non-TK districts is not particularly concerning. We observe several important and predictive characteristics that feed into our predicted 3rd grade math scores, and we find no discontinuity in this measure. If the density discontinuity reflects unobservable differences between students across the cutoff, these differences would have to be orthogonal to all the observable characteristics we account for, which seems highly unlikely.

Table A4: Covariate Continuity Through the Cutoff

	Control Mean	Estimate	Standard Error	P-value
Panel A. Non-TK Districts				
Predicted 3rd grade math score	−0.095	−0.015	0.019	0.434
Female	0.490	−0.006	0.019	0.755
White	0.491	−0.045	0.022	0.045
Black	0.355	0.022	0.020	0.256
Hispanic	0.096	0.011	0.010	0.254
Asian American	0.039	0.004	0.008	0.610
Other race	0.019	0.007	0.006	0.224
Economically disadvantaged	0.697	0.017	0.017	0.331
Prior state pre-K enrollment	0.000	0.088	0.007	0.000
Neighborhood White share	0.640	−0.012	0.014	0.396
Neighborhood poverty share	0.175	−0.004	0.008	0.580
Neighborhood unemployment rate	0.133	0.004	0.006	0.528
Neighborhood BA attainment rate	0.133	0.004	0.006	0.528
Neighborhood median HH income	49,213	−634	1,155	0.585
School is in a city	0.381	0.035	0.024	0.156
School is in a suburb	0.304	0.001	0.017	0.956
School is in a town	0.075	−0.024	0.012	0.046
School is in a rural area	0.240	−0.012	0.019	0.544
Magnet school	0.176	0.029	0.014	0.036
Log(school enrollment)	5.941	−0.007	0.016	0.692
School pupil:teacher ratio	17.8	0.0	0.2	0.806
School free- or reduced-price lunch share	0.668	0.005	0.012	0.717
Panel B. TK Districts				
Predicted 3rd grade math score	0.272	0.020	0.017	0.242
Female	0.511	−0.021	0.020	0.291
White	0.732	0.005	0.024	0.831
Black	0.124	−0.009	0.012	0.439
Hispanic	0.075	−0.004	0.012	0.711
Asian American	0.057	0.005	0.014	0.730
Other race	0.011	0.003	0.004	0.354
Economically disadvantaged	0.469	−0.026	0.019	0.166
Prior state pre-K enrollment	0.000	0.057	0.005	0.000
Neighborhood White share	0.849	0.005	0.008	0.521
Neighborhood poverty share	0.089	0.004	0.004	0.261
Neighborhood unemployment rate	0.197	0.001	0.004	0.863
Neighborhood BA attainment rate	0.197	0.001	0.004	0.863
Neighborhood median HH income	66,425	−11	1,099	0.992
School is in a city	0.190	0.009	0.015	0.540
School is in a suburb	0.505	0.005	0.019	0.812
School is in a town	0.122	0.001	0.016	0.940
School is in a rural area	0.184	−0.015	0.016	0.348
Magnet school	0.078	−0.006	0.012	0.594
Log(school enrollment)	6.056	−0.062	0.017	0.000
School pupil:teacher ratio	17.3	−0.4	0.1	0.018
School free- or reduced-price lunch share	0.429	0.007	0.008	0.437

D Robustness

D.A Models With and Without Covariates

Table A5 shows that our results are robust to the inclusion or exclusion of covariates in the impact models. The differences between the estimates are small and statistically insignificant.

In the preferred approach, we always exclude covariates when estimating EK LATEs because the demographic subgroups are defined by the covariates. Hence, the estimates of the EK LATEs are identical, by construction, for the with- and without-covariate results shown in the table. The TK LATEs, however, do change slightly when we exclude covariates from the estimation of the other pieces involved in backing out the TK LATE (i.e., ITT, Ω_{TK}, and Ω_{EK}). The estimates for math and ELA both increase slightly, but the differences are not statistically significant. Overall, the inclusion of covariates hardly matters for our estimates.

Table A5: 3rd Grade Test Score Impacts With and Without Covariates

	Math		ELA	
	Baseline	Preferred	Baseline	Preferred
	Panel A. With Covariates			
$LATE_{TK}$	0.212*	0.294	0.097	0.191
[P-value]	[0.051]	[0.111]	[0.401]	[0.293]
$LATE_{EK}$	−0.366***	−0.557*	−0.219*	−0.435
[P-value]	[0.000]	[0.092]	[0.078]	[0.181]
	Panel B. Without Covariates			
$LATE_{TK}$	0.252**	0.331*	0.123	0.209
[P-value]	[0.046]	[0.088]	[0.321]	[0.253]
$LATE_{EK}$	−0.378***	−0.557*	−0.240*	−0.435
[P-value]	[0.000]	[0.092]	[0.061]	[0.181]
Control mean	0.302		0.286	
Observations	15,680		15,669	

D.B Subgroup Estimates in the "Preferred Approach"

Table A6 shows the subgroup estimates and weights that feed into our "preferred" estimation approach. For each outcome domain, we estimate eight EK LATEs, one for each group defined by sex × race × economic disadvantage status. Most of the subgroup estimates are negative, as we expect, although some of the subgroups with small samples have imprecise positive estimates. We use the shares in the "Share in TK Districts" column as weights to aggregate the subgroup LATEs into a single $LATE_{EK}$ estimate. The shares in the "Share in Non-TK Districts" column are provided as a comparison point.

Our preferred estimate for $LATE_{EK}$ is greater in magnitude than our baseline estimate because demographic cells with large $LATE_{EK}$ estimates are a larger fraction of all EK compliers in TK

districts than in non-TK districts. In particular, female students who are White or Asian (regardless of economic disadvantage status) have large negative estimates and receive much more weight in TK districts than in non-TK districts. Recall from Table 1 in the paper that students in districts with TK are substantially more likely to be White.

Table A6: Subgroup EK LATE Estimates and EK Complier Shares

Sex	White or Asian	Economically Disadvantaged	$LATE_{EK}$	Share in TK Districts	Share in Non-TK Districts
			Panel A. Math		
Male	No	No	0.260	.02	.03
Male	No	Yes	−0.298	.09	.25
Male	Yes	No	−0.543	.15	.06
Male	Yes	Yes	0.362	.16	.09
Female	No	No	1.036	.03	.04
Female	No	Yes	−0.571	.12	.28
Female	Yes	No	−0.366	.22	.10
Female	Yes	Yes	−0.879	.21	.13
			Panel B. ELA		
Male	No	No	0.386	.02	.03
Male	No	Yes	−0.075	.09	.25
Male	Yes	No	−0.162	.15	.06
Male	Yes	Yes	0.237	.16	.09
Female	No	No	−1.558	.03	.04
Female	No	Yes	−0.208	.11	.28
Female	Yes	No	−0.107	.22	.10
Female	Yes	Yes	−0.638	.21	.13

Note: The point estimates in the $LATE_{EK}$ column are estimated using Equation 3 from the paper, but with covariates excluded because the subgroups are defined based on the covariates. The last two columns are the share of all EK students (within our 30 day bandwidth) in TK and non-TK districts who belong to each demographic cell. The shares do not always sum to 1 due to rounding. The shares are slightly different for math and ELA due to small differences in missing test score data by domain.

D.C Bounding the TK LATE

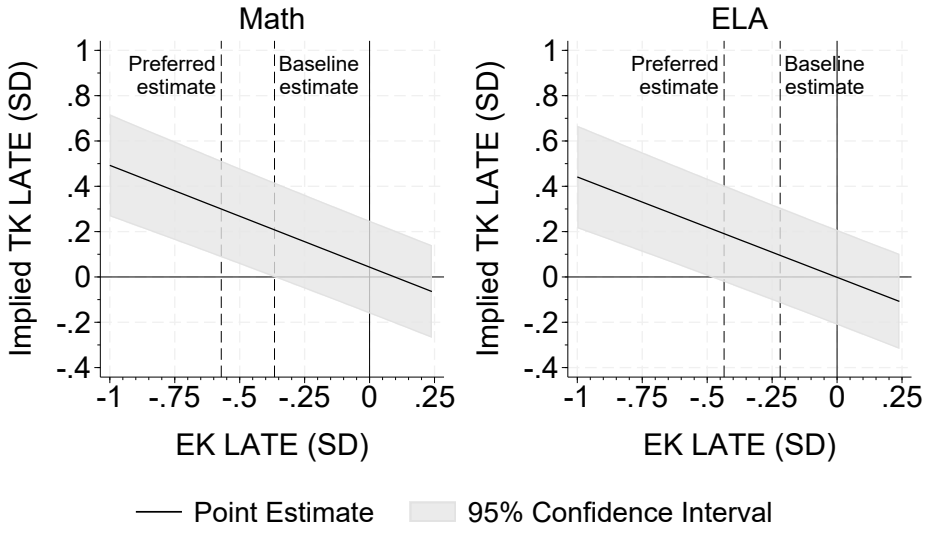
In the paper, we describe the subgroup-based approach as "preferred" because it relaxes the assumption regarding homogeneity of the EK treatment effect. Specifically, this approach assumes EK treatment effects may differ *across* the eight demographic groups, but *not* within demographic group across TK and non-TK districts. For example, it assumes the treatment effect of EK for white or Asian females who are not economically disadavantaged is identical in TK and non-TK districts.

In this section, we take a different approach to investigating the robustness of our TK LATE estimates. We explore how large or small the true EK LATEs would have to be to imply substantially different TK LATE estimates. Figure A2 shows the implied $LATE_{TK}$ estimate for every value of $LATE_{EK}$ ranging from −1.0 to 0.25 using Equation 1 from the paper and our primary estimates of ITT, Ω_{TK}, and Ω_{EK}.

For both outcome domains, we estimate a larger TK LATE when the EK LATE is larger. Because

the intent-to-treat discontinuity is slightly larger for math than ELA, the implied TK LATE is also slightly larger for math for any given EK LATE. For us to estimate that the TK LATEs for math and ELA are roughly 0, the true EK LATEs would have to be as small as 0.08 and −0.02 standard deviations, respectively. The math TK LATE would lose statistical significance at the 95% confidence level if the true EK LATE is less than −0.36 standard deviations, which is approximately our baseline EK LATE estimate. The EK LATE for ELA would have to be more negative than −0.5 standard deviations for us to estimate a statistically significant positive TK LATE for ELA.

Figure A2: Range of Possible TK Test Score Impacts



D.D Alternative Identification Assumptions

Disentangling the TK and EK treatment effects requires some restriction on treatment effect heterogeneity. For instance, our baseline approach assumes the EK LATE is the same in TK and non-TK districts; our preferred approach assumes the EK LATE is the same in TK and non-TK districts, but only within demographic groups. In this section, we discuss a third assumption that, in theory, would also allow us to disentangle the two treatment effects. In practice, however, this approach is uninformative for us.

The third potential assumption is that the EK impact is homogeneous across demographic groups, although it may differ across districts with and without TK. For example, with demographic groups defined based on sex, the assumption would be that boys and girls in TK districts have the same EK LATE. We view this assumption as complementary to our preferred approach; each assumption relaxes treatment effect homogeneity in one dimension while enforcing it in another. Our preferred approach relaxes homogeneity across student type, whereas this approach relaxes homogeneity across district type. Caetano et al. (2023) develop an identification argument and estimation techniques using this assumption.

Unfortunately, the third approach was not informative in our setting. Using the Caetano et al. (2023) estimator and only data from TK districts, in various specifications we defined demographic groups based on one of sex, race, economic disadvantage status, and cohort. The resulting estimates were too noisy for us to draw any conclusions.

Another student characteristic we considered using within the Caetano et al. (2023) framework was distance from one's neighborhood to the nearest in-district school that offers TK. As in other

settings, distance likely affects program take-up and plausibly does not separately influence academic outcomes. However, two issues prevented us from using distance with this approach. The first issue was power. In around half of all TK districts, every school with kindergarten students also has a TK program. In these districts, distance would not cause differential take-up of TK and EK. The second issue was the non-random placement of TK programs. In the other half of TK districts—the ones that offer TK in some but not every building with kindergarteners—TK programs are more likely to be in schools that serve more economically disadvantaged children. They are also more likely to be placed in elementary schools that have (non-TK) pre-K programs or whose highest grade is not higher than 3rd grade. These observable differences make it less likely that a homogeneous treatment effect assumption would hold between students who live different distances from TK programs.

E Inference

We conduct inference via bootstrap because our preferred approach to identification requires a multi-step estimation procedure. For consistency, we conduct bootstrap inference in our baseline approach too, although we get nearly identical results using the standard parametric approach.

Specifically, we implement a "Bayesian bootstrap" that creates new samples by reweighting rather than resampling. The procedure is stratified across TK and non-TK districts and clustered on the RD running variable (i.e., birthday). All observations within a cluster share a single replication weight, drawn randomly from an exponential distribution. The replication weights are normalized so that the sum of each cluster weight equals the number of clusters in a strata. We draw 1,000 sets of weights.

In the baseline approach, we estimate our two-stage least squares model with bootstrap weights 1,000 times. In the preferred approach, we re-estimate every part of the multi-step procedure 1,000 times. Doing a full bootstrap accounts for uncertainty in the 8 subgroup EK LATEs in non-TK districts; the 8 EK complier shares in TK districts; and the ITT, Ω_{TK}, and Ω_{EK} in TK districts.

In Table 2 of the paper we summarize our inference results using p-values rather than standard errors. We omit standard errors because they are uninformative in our preferred approach. The bootstrap distributions are highly non-normal in our preferred approach, containing some extreme outliers. These outliers likely exist because we split the sample into small subgroups and estimate a large number of parameters, which creates several opportunities for sampling variation to produce extreme outcomes. Consequently, the outliers drive up the TK and EK LATE standard errors, making them uninformative about variation throughout most of the distributions.

Instead of standard errors, Table 2 presents p-values from two-tailed hypothesis tests. We calculate p-values in two steps. First, for each estimate, we enforce a null hypothesis that there is no effect by subtracting the mean of the bootstrap distribution from each bootstrap estimate. Second, we calculate the share of demeaned estimates that are greater (in absolute value) than our primary point estimate. This share is the p-value.

References

Caetano, Carolina, Gregorio Caetano, and Juan Carlos Escanciano, "Regression Discontinuity Design with Multivalued Treatments," *Journal of Applied Econometrics*, 2023, *38* (6), 840–856.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma, "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 2020, *115* (531), 1449–1455.

McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 2008, *142* (2), 698–714.

Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, Pam Broene, Frank Jenkins, Andrew Mashburn, and Jason Downer, "Third Grade Follow-Up to the Head Start Impact Study: Final Report," Technical Report OPRE Report 2012-45, Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, Washington, DC 2012.