Kumon Up: Private Tutoring Centers and Student Achievement

Joshua Goodman
Boston University
edecjosh@bu.edu

Samuel Nitkin
Boston University
snitkin@bu.edu

July 16, 2026*

Abstract

Private tutoring chains like Kumon and Sylvan have expanded rapidly in recent decades. Combining data on business locations, household spending, and district test scores, we study who uses these centers and how they affect academic skills. Nearly half of users are from the top income decile, with Asian and White households five and two times as likely to purchase tutoring as Black and Hispanic households. A staggered differences-in-differences design shows the opening of a district's first chain center raises district-wide math and reading scores, at a lower cost than public spending typically producing similar gains. Math effects are driven by White and Asian students, so that center entry widens racial achievement gaps, though less so in more racially integrated districts. These private investments help their users, deepen educational inequality, and shape outcomes by which state accountability systems and families judge schools.

*For their helpful feedback, we thank seminar participants at NBER Summer Institute, Brown University, Boston University, and DIW Berlin.

1 Introduction

Private tutoring centers have grown into a common feature of K-12 education in the United States. Chains like Kumon, Sylvan, and Mathnasium have expanded from a few hundred U.S. locations in the late 1990s to several thousand today, and millions of families now pay for instruction outside the regular school day. Despite this growth, remarkably little is known about who uses these centers or whether they improve student achievement. This gap matters for at least two reasons. First, a growing body of evidence suggests that intensive, individualized tutoring can substantially raise student achievement, making private tutoring centers a potentially important and scalable complement to or substitute for in-school instruction. Second, if access to this kind of instruction is unevenly distributed, private tutoring centers could widen existing gaps in family investment in children. This paper provides the first comprehensive evidence on private tutoring center usage and impact across the United States. We ask two questions: Who uses chain tutoring centers? And what effect does their entry into a community have on student achievement?

We begin with descriptive analysis of tutoring centers and the families who use them. Business registry records from Data Axle provide 25 years of tutoring center locations across the United States. We focus on national chains, which are easiest to identify and about which we best understand the treatment itself. The business registry data show that chain tutoring centers have grown substantially since the late 1990s, expanding from roughly 500 locations in 1997 to a peak of over 4,000 by the mid-2010s. Of the over 7,000 unique center locations during our main period of interest, Kumon accounts for 38 percent, Sylvan Learning 25 percent, Mathnasium 10 percent, and Huntington Learning 10 percent, together representing over four-fifths of all chain locations in our data. A few chains focus solely on math instruction, and math appears to dominate the business model for chains offering both math and reading support. The typical center likely serves 100 to 130 students at a time, with the typical students spending two to three hours per week at the center or doing center-assigned homework.

The Consumer Expenditure Survey (CEX) allows us to understand who uses tutoring centers by measuring household spending on tutoring by geography, race, and income. Household spending data suggests tutoring usage is rare overall but sharply stratified by income and race. Just 1.7 percent of households with school-age children report any monthly tutoring spending, averaging about $300 among spenders. Asian households are far more likely to use tutoring than any other group, with 5.2 percent spending on tutoring, more than triple the national rate. The fractions of White, Black, and Hispanic households who spend on tutoring are 1.8, 1.2, and 0.8 percent respectively. This racial gap is driven largely but not entirely by income. Top-decile households account for 47 percent of households who report spending on tutoring. Among top-decile households, White, Black, and Hispanic usage is roughly 4-5 percent, while Asian usage remains far higher at nearly 10 percent. Conditional on income at all levels, White, Black and Hispanic households have roughly similar spending patterns while Asian families are more likely to spend on tutoring. Geographically linking spending to local tutoring chain density shows that tutoring expenditures track the availability of physical chain centers, but only among White, Asian and above median income households. This suggests that chain centers in particular, rather than tutoring broadly, are disproportionately used by these groups.

To estimate the impact of tutoring centers on student outcomes, we link center locations to test score data and implement a staggered difference-in-difference approach. Test score data come from the Stanford Education Data Archive (SEDA), which provides standardized test score outcomes for nearly all school districts in the country, including by demographic subgroups. We match each tutoring center location to a school district in SEDA and merge center counts in a given calendar year with test scores from the following spring. We then exploit variation in the timing of a district's first chain center opening, comparing districts that gain a first chain between 2010 and 2019 to districts that never have a chain center. Because staggered treatment timing can generate biased estimates under standard two-way fixed effects approaches with heterogenous treatment effects, we use the two-stage difference-in-differences estimator of Gardner et al. (2024). The first stage estimates district-by-grade and year fixed effects using only never-treated and not-yet-treated observations, and the second stage regresses each observation's residual outcome (net of these fixed effects) on event-time indicators. This approach yields estimates of chain centers at each year relative to first opening that are not contaminated by comparisons to already-treated units.

Chain tutoring centers have a strong, positive, and persistent impact on student achievement. Following the opening of a district's first chain center, overall math test scores rise by 0.029 standard deviations and reading scores by 0.016 standard deviations, with both effects growing over the five years following entry. Event-study estimates show flat pre-trends for both subjects, supporting the parallel trends assumption underlying our design. These results are robust to controlling for grade-level enrollment and racial composition and for district-level unemployment and median income, with controlled estimates of 0.034 standard deviations for math and 0.021 for reading. The relative size of the math and reading effects is roughly proportional to our best evidence on the fraction of students using tutoring centers for those subjects, suggesting that such differences may be accounted for by differences in implied first stage impacts on usage rates.

This overall effect masks substantial racial heterogeneity and is driven entirely by White and Asian students. Math scores for White and Asian students rise by 0.040 standard deviations following chain entry, while effects for Black, Hispanic, and Native American students are statistically indistinguishable from zero. Consistent with this divergence, chain entry widens within-district test score gaps, increasing the White-Black math gap by 0.037 standard deviations and the White-Hispanic gap by 0.020 standard deviations. These gap widening effects are substantially larger in more racially segregated districts (0.047 standard deviations for the White-Black gap in highly segregated districts versus 0.021 in less segregated ones), consistent with a within-school spillover mechanism in which the benefits of tutoring also accrue to classmates less likely to be using tutoring centers.

In contrast to this racial heterogeneity, chain tutoring's effects show little variation along other dimensions. Math score gains are similar for economically disadvantaged and non-disadvantaged students and the within-district gap between these groups shows no significant change following chain entry. Effects are likewise comparable for male and female students, and only modestly larger for younger students in grades 3-5 than for those in grades 6-8. Non-chain tutoring centers and chains that provide in-home tutors show statistically insignificant effects on math and reading test scores across subgroups, suggesting that the achievement gains we document are specific to brick-and-mortar chain centers rather than a general feature of private tutoring.

Our headline estimates are robust to alternative empirical choices. Replacing the two-stage difference-in-differences estimator with standard two-way fixed effects yields a similar effect on math scores, and replacing the binary first-chain indicator with a continuous measure of chain centers per 1,000 children—extended to all school districts rather than just our first/never-chain sample—continues to show positive, significant effects on math achievement. Results are similarly robust to an alternative geographic definition of treatment, holding when we define district exposure using chain centers within a 10-mile radius of the district centroid rather than within district boundaries.

How large are these effects for the students who actually attend these centers? Because our district-level estimates average over the many students who never enroll, the implied effect on tutored students is necessarily larger. Scaling our intent-to-treat estimates by the share of students induced to take up tutoring—inferred either from household expenditure patterns or from the physical capacity of the centers themselves yields a per-student effect between roughly 0.3 and 1.0 standard deviations. This range straddles the 0.4 standard deviation gain produced by the average tutoring program in meta-analyses of the experimental literature, suggesting our estimates are reasonable(Nickow et al., 2024). The gains also appear strikingly cost-effective: the tutoring that generates our district-average effect costs on the order of $70 to $90 per pupil, roughly an order of magnitude less than the increase in public school spending that Jackson and Mackevicius (2024) find produces a comparable test-score gain, albeit borne privately by families rather than by school systems.

These effects are also large enough to account for much of how treated districts' test scores actually evolved over our sample period. Between 2009 and 2019, math scores in these districts rose by 0.046 standard deviations, so that our estimated center-entry effect of 0.029 explains nearly two-thirds of their improvement; the analogous figure for reading is just over half. A similar exercise attributes about half of White and Asian students' gains to chain entry, and a substantial share of the within-district widening of racial achievement gaps, including more than 40 percent of the increase in the White-Black gap. Private investments made outside the public system thus appear to have non-trivial effects on the outcomes by which districts are judged by families and state accountability systems.

This paper relates to three strands of research. First and most narrowly, it contributes to a small literature documenting the expansion and educational impacts of private tutoring centers. Kim et al. (2025) documents the rapid recent expansion of U.S. tutoring centers, showing that they tend to locate in higher-income, heavily Asian-American areas. We corroborate this pattern with household spending data and build on it using a causal design. The cleanest causal evidence on the efficacy of tutoring center curricula comes from a randomized controlled trial conducted in Bangladeshi primary schools, where Kumon directly provided instructional materials matching those provided in its U.S. centers. Sawada et al. (2024) find that this self-directed, mastery-based approach generated large gains in math achievement among participants, while a companion paper, Kawarazaki et al. (2023), showed positive peer effects. Most closely related to our paper is contemporaneous work by Chen (2026), who studies the expansion of private tutoring centers in Texas. He finds that access to such centers raises SAT and ACT scores, shifts students toward more selective and out-of-state colleges, and increases early-career earnings. The effects are entirely concentrated among non-economically disadvantaged students, who his original survey evidence suggests are more likely to take up tutoring. Whereas Chen (2026) focuses on one large state, college entrance exams, and postsecondary outcomes, our paper complements this evidence by studying the national landscape of chain tutoring centers and their impact on primary grades academic skills, including those measured by state standardized exams unlikely to be a particular focus of tutoring center curricula.

Our paper also speaks to the broader literature on effective approaches to improving academic achievement, including other forms of tutoring. Tutoring centers' efficacy is likely due to some combination of the individualized instruction they provide and the increased instructional time they entail for students. Meta-analyses document the particular effectiveness of small group and one-on-one tutoring (Nickow et al., 2024; Fryer Jr, 2017), a finding exemplified by recent experimental evaluations (Escueta et al., 2020; Carlana and La Ferrara, 2025; Guryan et al., 2023; Kraft et al., 2022). Computer-assisted learning has also garnered attention for its positive impacts (Muralidharan et al., 2019; Bhatt et al., 2024; Oreopoulos et al., 2024), largely attributed to the ability to individualize learning to students. Evidence that instructional time is also a key input into student achievement comes from quasi-experimental variation in individual absences (Aucejo and Romano, 2016; Liu et al., 2021; Goodman, 2014), policy-driven changes to daily or weekly school calendars (Figlio et al., 2018; Thompson, 2021), descriptive evidence from international achievement comparisons (Lavy, 2015; Rivkin and Schiman, 2015), and both descriptive and experimental evidence showing that longer school days and years may partly explain charter school efficacy relative to traditional public schools (Angrist et al., 2013; Fryer, 2014). Kraft and Novicoff (2024) conclude from this literature that the impact of one additional hour of instruction per week in a specific subject is roughly 0.02-0.07 standard deviations. Even doubling or tripling this figure to account for the typical weekly time students spend at tutoring centers yields a smaller effect than implied by our estimates, suggesting that tutoring centers may be particularly effective relative to typical instructional time.

Finally, this paper addresses the growing inequality in child development driven by unequal family investment. Tutoring centers are one example of the increasing temporal and financial investments high income families are making in their children. Ramey and Ramey (2010) argue that competition for scarce college slots generates a “rug rat race”, explaining why time spent by parents on childcare in the U.S. began rising dramatically in the mid-1990s, particularly among college-educated parents. Doepke and Zilibotti (2019) argue that increasing inequality in some countries, such as the United States, drives parents to push their children harder to ensure their success, transforming the hands-off parenting of the 1960s and ’70s into something more like helicopter parenting with a full schedule of after-school activities. Kornrich and Furstenberg (2013) show that, from 1972 to 2005, the gap between top and bottom income quintile families’ spending on child-enrichment goods and services nearly tripled, from roughly $2,700 to $7,500 per child per year. Schneider et al. (2018) shows that class gaps in parental financial investments are wider when state-level income inequality is higher, while spending on things other than children shows no such pattern. Blazar et al. (2026) provide the most comprehensive recent accounting, finding that average per-child investments from age 5 to 18 across all domains are 12 percent lower for bottom vs. top income quartile households, 13 percent lower for Hispanic vs. White children, and 7 percent lower for Black vs. White children. Our paper contributes to this literature by causally identifying private tutoring centers as one concrete channel through which these investment gaps translate into widening academic achievement gaps.

2 Tutoring Centers

Private tutoring centers in the United States operate under a variety of instructional models but share a common structure as for-profit supplemental education providers serving K–12 students outside of school hours. We focus our discussion on the chains most prevalent in our data: Kumon, Sylvan Learning, Mathnasium, and Huntington Learning Center.

Kumon, founded in Japan in 1954 and the largest chain in the sample by a substantial margin, operates on a self-directed mastery model in which students work independently through printed worksheets calibrated to their current level, advancing only upon demonstrating both accuracy and speed. Instructors in a Kumon center do not teach in the conventional sense. They circulate, grade, and monitor, while the structured worksheet sequence serves as the primary vehicle of instruction. Students are also expected to complete daily homework packets at home, making consistent parental involvement a prerequisite for the program's intended effect. Parents must both transport their child to a center at least once weekly and supervise daily at-home practice on the remaining days. Kumon offers both math and reading curricula.

Sylvan Learning, the next largest chain, uses a more instructor-focused approach. It employs certified instructors who deliver direct instruction through a proprietary technology-assisted curriculum, with sessions conducted in small groups of two to three students. Sylvan's curriculum is deliberately aligned to local school standards, and instructors are encouraged to communicate with students' classroom teachers, positioning Sylvan less as an independent enrichment program and more as a school-adjacent intervention. Like Kumon, Sylvan begins with a diagnostic assessment and constructs an individualized learning plan, but the mechanism of instruction differs. Kumon relies on the student to work through material independently while Sylvan relies on the instructor to actively teach it. Parental involvement at Sylvan is more limited than at Kumon, as the program does not assign daily homework packets. Parents are primarily responsible for transporting their child to center sessions two to three times per week and attending periodic progress meetings with the center director. Sylvan offers both math and reading curricula.

Mathnasium focuses exclusively on math, with a proprietary instructional method delivered through verbal, visual, and written exercises by instructors trained in the "Mathnasium Method." Like Sylvan, it operates physical centers and serves multiple students simultaneously, but each student follows an individualized plan rather than a shared curriculum sequence. Mathnasium explicitly discourages students from bringing school homework to sessions, positioning itself as a standalone math curriculum rather than a supplement to classroom assignments. Parental involvement is limited to transportation and occasional progress check-ins. Huntington Learning Center, one of the oldest chains in the sample, similarly offers center-based tutoring and is notable for having sought formal accreditation for its locations, an unusual step in the supplemental education sector that signals a closer alignment with institutional schooling norms than most competitors pursue.

Despite their differences, these chains share several structural features relevant to understanding their potential effects on student achievement. All administer diagnostic assessments at intake and construct individualized learning plans; all focus primarily or exclusively on math and reading, the subjects measured in our outcome data; and all operate as franchise systems in which local owners hire and manage instructors under a corporate curriculum and brand. The standardized curricula, instructor training requirements, and quality oversight mechanisms embedded in franchise agreements may partly explain why chain centers produce measurable achievement gains while independent tutoring centers do not.

Franchise disclosure documents give us additional insight into the number of students using each center and the ways in which they are typically used.[1] Our best estimate suggests the typical tutoring center in our data serves 100 to 130 students at a given time, though the underlying calculation differs by chain depending on what each disclosure document reports.

For Kumon, whose franchise royalty is a fixed per-student fee, we divide total royalty revenue reported in Kumon's financial statements by this per-student rate and by the number of U.S. centers, yielding an estimate of roughly 150 subject-enrollments. Assuming 15 to 35 percent of students are dual-enrolled in math and reading, this implies 100 to 130 distinct students per center.[2] For Huntington, whose franchise disclosure document reports both average hourly tuition and average annual hours of instruction per student by service line, we instead divide reported per-center revenue by the implied revenue per student (tuition rate multiplied by hours of instruction), yielding a similar estimate of 110 to 130 students per center per year.[3]

Figures are less certain for Sylvan and Mathnasium, both of which charge royalties as a percentage of gross revenue, so no student count can be recovered from the royalty structure itself. Sylvan's document contains no comparable assumption; instead, we infer that centers typically enroll close to 100 students from a staffing rule requiring any center with more than 100 enrolled

[1]We obtained 2024 franchise disclosure documents from the Minnesota Department of Commerce website at `https://cards.web.commerce.state.mn.us/franchise-registrations`.

[2]Another data point comes from a 2014 Kumon press release claiming its 1,492 centers served 265,000 U.S. students, implying about 178 students per center. See the press release at: `https://www.kumonfranchise.com/us-en/assets/pressrelease/release2014-0820.html`. That figure likely includes students using Kumon for only part of the year, thus explaining why it is higher than the franchise-disclosure-document calculation, which assumes full-year participation.

[3]The 110 to 130 range for Huntington reflects that only Learning Center and Test Prep enrollment (about 68 and 39 students, respectively, for a total of 107 students) can be calculated directly from disclosed hourly tuition and hours-per-student. The remaining 14 percent of revenue from Subject Tutoring and Other services, for which no price or hours data are disclosed, is assigned an estimated 20 or so students.

students to hire a second director. For Mathnasium, we rely on the company's own hypothetical reference case: its franchise disclosure document sets required insurance coverage based on "an assumed enrollment of 100 students" per center. Neither figure is derived from reported revenue and should be treated as an order-of-magnitude benchmark rather than a calculated estimate.

The best evidence we have suggests that math is a larger focus of these centers than reading. For Kumon, two pieces of evidence suggest 60 to 75 percent of Kumon tutoring is focused on math. First, to become a fully licensed franchisee, a Kumon center must cumulatively enroll 600 math students and 400 reading students, suggesting 60 percent of its focus is math. Second, we spoke to a Kumon franchising executive, who suggests that 70 to 75 percent of U.S. Kumon tutoring is in math, with the remaining 25 to 30 percent in reading. This suggests that two to three times as many students use Kumon for math as for reading support. We know little about subject usage rates at Sylvan and Huntington. Mathnasium only provides math tutoring.

The typical tutored student appears to spend roughly two to three hours per week in these centers or doing center-assigned homework. Franchise disclosure documents give a partial picture of weekly instructional intensity. Kumon states that students attend the center twice weekly for 20 to 30 minutes per subject, and additionally assigns homework on the five non-Center days each week. Assuming 20 minutes of daily homework per subject, consistent with the length of a typical Kumon worksheet set, a single-subject Kumon student spends roughly 50 minutes in-center plus 100 minutes on homework, for a total of about 2.5 hours per week. Mathnasium, by contrast, discloses that students typically attend twice weekly for approximately 60 minutes per session, or about 2 hours per week, with no homework component. Sylvan and Huntington do not disclose comparable attendance figures but their marketing materials suggest a similar order of magnitude to Mathnasium.[4]

[4]Sylvan's website and individual center pages recommend two to three one-hour sessions per week (`https://www.sylvanlearning.com/how-sylvan-works/faqs/`), implying roughly 2–3 hours per week with no structured homework component. Huntington's website recommends at least two sessions per week during the school year (`https://huntingtonhelps.com/why-huntington/`), and states explicitly that homework is not assigned for general academic tutoring (`https://huntingtonhelps.com/resources/faqs/`). At roughly one hour per session, this implies about 2 hours per week during the school year.

3 Data

Our descriptive analysis primarily relies on historical data on tutoring center locations, from Data Axle, and household spending data, from the Consumer Expenditure Survey. Our causal analysis relies on the tutoring center location data combined with data on student test scores, from the Stanford Education Data Archive. We describe these three main data sources in further detail.

3.1 Tutoring Centers

We measure the locations of private tutoring centers using Data Axle's Historical Business Data, which collects information on businesses sourced and verified by yellow page directories, credit card billing data, phone verification, web research, news publications, and annual reports (Kim et al., 2025). We obtain records from 1997-2022, observing businesses' names, addresses, and industry codes. Each business included in the data represents a unique physical location, with identifiers that allow the business to be tracked across years.

We define tutoring centers as businesses labeled either "Tutoring" (SIC Code 829909) or "Test Preparation Instruction" (SIC Code 874868). We exclude the 5 percent of observations where the firm name clearly indicates a focus on college test-preparation, graduate, ESL, or non-academic (e.g. music) tutoring. We define chains as 10 or more unique locations that share a parent company (as identified by Data Axle) or a common name. We identify those common names primarily through our own inspection, supplemented with the use of Claude Sonnet 4.6 to catch repeated name patterns that we missed. We further restrict our main analysis to chains that provide in-center tutoring; we analyze chains that provide in-home tutors in a later sub-analysis. Of the initial set of 20,337 unique tutoring locations, 35 percent (7,071) are identified as chains providing in-center tutoring services.

The business registry data show that chain tutoring centers have grown substantially since the late 1990s. As Figure 1 shows, the number of such centers grew from roughly 500 locations in 1997 to a peak of over 4,000 by the mid-2010s, plateauing in more recent years. Kumon is by far the largest of these chains and has grown the most over this time, with Sylvan Learning, Mathnasium, and Huntington Learning being the next largest chains. Table 1 provides more details on these chains during the 2009-19 period for which we have test score data. Of the over 7,000 unique center locations during that period, Kumon accounts for 38 percent, Sylvan Learning 25 percent, Mathnasium 10 percent, and Huntington Learning 10 percent. Those four chains together represent over four-fifths of all chain locations, with 19 other identified chains comprising the remainder. All of the largest chains but Mathnasium offer both math and reading instruction, with some smaller chains specializing in either math or reading.

3.2 Consumer Expenditure Survey

We measure household-level consumption of tutoring with the Consumer Expenditure Survey (CEX), a nationally representative survey conducted by the U.S. Census Bureau that tracks household spending and characteristics. Beginning in 2009, the CEX includes a dedicated expenditure category for "Test preparation or tutoring services." Using CEX public-use microdata from January 2009 to February 2024 (U.S. Bureau of Labor Statistics, 2024), we construct two measures of monthly tutoring usage: (1) a binary indicator for whether the household reports any spending on tutoring services in a month, and (2) the amount spent on tutoring services, inflation-adjusted to 2015 dollars using the Consumer Price Index for All Urban Consumers (CPI-U). We note that these measures capture expenditures on all forms of tutoring (in-center, in-home, online), not just tutoring at the physical locations of chain centers, which are the focus of this study. We further record the race/ethnicity of the household reference person and the household's income percentile. To align measures of tutoring expenditure with academic-year outcomes, we restrict the CEX to households with at least one child ages 6-17 and exclude spending reported in June, July, and August. This restriction yields a final sample of 173,328 household-month observations representing 70,590 unique households.

3.3 School District Test Scores

Our primary outcomes are district-level math and reading standardized test scores from the Stanford Education Data Archive (SEDA) (Reardon et al., 2024). SEDA provides mean test scores by school district for grades 3–8 over the period 2009–2019, including breakdowns by racial and ethnic subgroup, gender, and economic disadvantage status.[5] Scores are reported using SEDA's

[5] We follow SEDA convention and refer to academic years by their spring term, so that 2009 refers to the 2008–09 academic year.

Cohort Scale, which enables consistent comparisons within a grade across districts and over time, expressed in standard deviations of the national distribution.

We match each tutoring center to a school district using 2019 district boundary shapefiles from SEDA, yielding a consistent geographic panel. Across all tutoring centers in the United States, 99.9 percent are successfully matched to a school district, with the remainder either lacking precise location data or falling outside any district boundary. We merge center counts in a given calendar year with test scores from the spring of that academic year. Center counts in 2014, for example, are matched to spring 2014 scores (i.e. from the 2013-14 school year). Because we observe only the calendar year of tutoring center entry rather than the precise date, we are uncertain whether a center opening in a given year was available to students prior to standardized testing. The year of center entry thus has ambiguous treatment status, whereas those before and after the entry year can be cleanly categorized as pre- and post-test administration. SEDA also provides school district characteristics that we later use as controls in some specifications, including total enrollments by grade, the racial composition of the test-taking population, the local unemployment rate, and the district's median household income.

4 Who Uses Tutoring Services?

Before causally estimating the impacts of tutoring centers on student achievement, we describe tutoring consumption in the United States. Among households with school-age children, 1.7 percent report spending on tutoring in a given month, as seen in Figure 2. Households that spend on tutoring spend on average $289 per month.[6] There is heterogeneity in tutoring use by race, with usage concentrated among Asian and, to a lesser extent, White families. The leftmost columns in Figure 2 show that 5.2 percent of Asian households spend on tutoring services each month, more than three times the national average. In comparison, 1.8 percent of White households consume tutoring services each month, while only 1.2 percent of Black households and 0.8 percent of Hispanic households do so.

High income households are much more likely to spend money on tutoring services than other households. As shown in the second group of columns in Figure 2, among households in the top

[6] See Panel A of Appendix FigureA.1 for details.

10 percent of the income distribution, 5.0 percent consume tutoring in any given month, compared to 1.5 percent in the next 40 percent of the income distribution and 0.6 percent of households in the bottom half of the income distribution.[7] Nearly half (47 percent) of all households that spend money on tutoring are in the top income decile.

Racial differences in spending are largely explained by differences in income. Among households in the top 10% of the income distribution, a similar share of White, Black, and Hispanic households consume tutoring services, and such shares are also relatively similar in other parts of the income distribution. Asian households are the exception, with tutoring take-up rates more than twice as high as any other group even conditional on income. These findings are consistent with geographic patterns documented by Kim et al. (2025), who find tutoring centers are concentrated in high-income and heavily Asian-American areas, and by family survey data from Chen (2026), who finds White, Asian and high income families more likely to report tutoring usage.

The CEX data captures all forms of tutoring, not just tutoring at chain centers, so we explore whether tutoring expenditures in that data correlate with geographic availability of chain centers. For each Metropolitan Statistical Area (MSA) in each year, we calculate the number of tutoring chain centers per 1,000 children ages 5-17.[8] As seen in Figure 3, binned scatterplots of household tutoring participation against MSA-year tutoring chain center density show strong positive relationships between the number of chains per 1,000 children and the share of households using tutoring services, indicating that our measure of tutoring participation in the CEX likely captures, at least partly, the use of physical tutoring centers. This relationship is particularly pronounced among White and Asian households and above median income households, but is statistically insignificant for other households. That the correlation between tutoring consumption and private tutoring center availability is driven almost entirely by White, Asian and higher income households suggests these are the families largely using tutoring centers.

[7]See Panel B of Appendix FigureA.1 for details on tutoring spending across the full income distribution.

[8]MSAs, clusters of counties with high population density, are the finest level of geography available in the CEX, and CEX sampling is stratified at the MSA level. Prior to 2016, geography is available for households in the 21 largest MSAs. In 2016 and on, the 23 largest MSAs are available. Overall, 45% of households in the CEX can be matched to an MSA.

5 Causal Sample and Method

5.1 Sample Construction and Summary Statistics

We now turn our attention to estimating the impact of tutoring center entry on student achievement. Our primary analysis sample is the panel of U.S. public school districts from 2009 to 2019 for which we observe test score data in at least 10 of those 11 years. We observe test scores and demographic characteristics for students in grades 3 through 8. Column (1) of Table 2 reports summary statistics for all districts at the beginning of the sample period. In 2009, we observe over 40,000 unique district-grade observations across nearly 9,000 districts. The average district has just under 5,000 children aged 5 to 17 and enrolls 344 students per grade. Test scores are normalized relative to the national distribution, so the average math and reading scores across all districts are very close to the national mean. In the typical district, 75 percent of students are White, 2 percent are Asian, 44 percent are economically disadvantaged, and 23 percent attend suburban schools.

Our main empirical strategy estimates the effects of a private tutoring chain opening in a district for the first time. For each district, we identify the first school year in which a chain center operated within district boundaries.[9] We denote treated districts, the 400 or so that received their first chain center between 2010 and 2019, as "First Chain" districts. Control districts are "Never Chain", the nearly 7,000 that never had a chain center between 1997 and 2019.[10] There are 1,924 first-chain district-grade observations across 402 districts and 33,758 never-chain district-grade observations across 6,959 districts.

Treated (First Chain) districts are larger, higher scoring, more Asian, higher income, and more suburban than Never Chain districts. Columns (2) and (3) of Table 2 summarize baseline characteristics for First Chain and Never Chain districts. First Chain districts' average math and reading scores are roughly 0.17 to 0.18 standard deviations higher than the national average, suggesting tutoring chains tend to open in already high-achieving areas. Those districts are also higher income, as seen by their lower rate of economic disadvantage and their higher median income. Asian students represent nearly 5 percent of First Chain districts' populations, compared to 1-2 percent

[9] Our tutoring center data begin in 1997. If a center opened and closed in a district prior to 1997, we would misclassify a later opening as the first opening. Our test score data begins in 2009, and we are generally unconcerned with the effects of centers that closed more than twelve years before the outcomes panel begins.

[10] This design excludes many large urban districts, which already had private tutoring centers prior to 2010. In section 6.3, we use an alternative empirical strategy to include these large urban districts.

in all districts or in Never Chain districts specifically. Over half of students in First Chain districts attend suburban schools, more than twice the fraction in comparison districts. These patterns are consistent with Kim et al. (2025), who show that private tutoring centers tend to locate in higher-income areas.

5.2 Empirical Strategy

To causally estimate the effect of tutoring centers on student test scores, we use a staggered event-study design driven by the first opening of a tutoring chain in a school district. There is large variation in the timing of tutoring chain entry, which allows us to compare changes in test scores within districts before and after center entry while absorbing common year shocks.[11]

Districts that receive a first center typically continue to have centers operating, but do not see a further increase in the number of private tutoring centers. As seen in Figure 4, in the five years after first entry, the average treated district has roughly 0.8-0.9 tutoring centers operating within its boundaries, consistent with the vast majority of new centers remaining open but not attracting competition from other centers. Scaled by the local population, first chain entry increases the number of centers per 1,000 school-aged children by 0.271 over the five years following first entry. This will be helpful later for understanding the magnitudes of the achievement effects we observe. Both measures of exposure imply that tutoring chain entry predicts sustained exposure to tutoring centers, and such magnitudes can be interpreted as a quasi-first stage for the stock of private tutoring in a district.

A standard two-way fixed effects (TWFE) event-study regression is problematic in this setting because of staggered treatment timing (Goodman-Bacon, 2021; Sun and Abraham, 2021). With staggered adoption, already-treated districts serve as controls for newly treated districts, and when treatment effects vary across units or over time, the resulting TWFE coefficients may place negative or otherwise un-intuitive weights on some comparisons. To address these concerns, we use the two-stage difference-in-differences event-study estimator proposed by Gardner et al. (2024). The estimator first estimates untreated potential outcomes using time and unit fixed effects from never-treated and not-yet-treated units, and then compares treated observations to their estimated untreated counterfactual. In the first stage, we estimate

[11] See Appendix Figure A.2 for more details.

$$Y_{dgt} = \alpha_{dg} + \lambda_t + X_{dgt}\beta + \epsilon_{dgt} \tag{1}$$

using only never-treated and not-yet-treated district-grade-year observations. Here, d indexes school districts, g indexes grade levels, and t indexes years. Y_{dgt} is the mean test score in district d, grade g, and year t; α_{dg} are district-by-grade fixed effects; λ_t are year fixed effects; and X_{dgt} is a vector of time-varying district-level characteristics. Each district-grade observation is given equal weight in this specification.

We then use the first-stage estimates to construct an adjusted outcome for every observation in the sample,

$$\widehat{Y}_{dgt} = Y_{dgt} - \widehat{\alpha}_{dg} - \widehat{\lambda}_t - X_{dgt}\widehat{\beta}. \tag{2}$$

This adjusted outcome removes the district-grade fixed component, common year shocks, and estimated effect of time-varying characteristics as recovered from untreated observations. In the second stage, we regress the adjusted outcome on event-time indicators:

$$\widehat{Y}_{dgt} = \sum_k \delta_k D^k_{dgt} + u_{dgt}, \tag{3}$$

where D^k_{dgt} is an indicator equal to one if district-grade dg in year t is k years relative to the opening of its first tutoring chain. Because the unit and time fixed effects are absorbed in the first stage rather than the second, the second-stage regression does not difference out a within-unit normalization, and all event-time indicators are separately identified; there is no mechanically omitted period. Standard errors are obtained by bootstrapping both stages and clustering at the district level. The bootstrap is essential here: because the second-stage outcome depends on first-stage estimates, it corrects the standard errors for the generated-regressor problem induced by the two-step procedure.

The coefficients δ_k have a natural event-study interpretation. For post-treatment periods, δ_k estimates the average effect on district-wide test scores k years after the first tutoring center opens, among district-grade-year observations observed at that relative time. These event-study coefficients are intent-to-treat (ITT) effects, measuring the change in average test scores across all students in a treated district, rather than treatment-on-treated (TOT) effects on the relatively small fraction of students who use tutoring centers. These effects are measured relative to the untreated counterfactual implied by the first-stage fixed effects and controls. For pre-treatment periods, the coefficients are placebo estimates: under parallel trends, they should be close to zero. In our main exhibits, we present the coefficients δ_k in event study figures.

We aggregate these event study coefficients to produce a single estimate of the effect of tutoring centers on test scores. Let $\#_{Pre}$ and $\#_{Post}$ denote the number of pre- and post-treatment event-study periods in the baseline specification; in practice, we set both equal to five. Our main summary estimand is the average post-treatment event-study coefficient minus the average pre-treatment coefficient:

$$ATT = \frac{1}{\#_{Post}} \sum_{k=1}^{\#_{Post}} \delta_k - \frac{1}{\#_{Pre}} \sum_{k=-\#_{Pre}}^{-1} \delta_k. \tag{4}$$

This estimand summarizes the change in adjusted test scores after treatment relative to the average pre-treatment placebo estimate. Rather than normalizing to a single reference period such as $k = -1$, we subtract the average of all pre-treatment coefficients, which differences out any residual level shift in the pre-period and avoids sensitivity to the choice of a single base year. We exclude $k = 0$ from the post-treatment average because center openings may occur at different points within the school year, making exposure during the opening year partial and difficult to interpret.

Our causal interpretation relies on the parallel trends assumption: absent the opening of a tutoring center, test scores in treated districts would have followed the same trends as test scores in control districts. The most natural threat to this assumption is that chains select into districts on the basis of anticipated test-score gains, so that treated districts were already on differential trajectories at the time of entry. We provide evidence against this concern by confirming that pre-treatment coefficients are statistically indistinguishable from zero and exhibit no systematic trends. We note that this evidence speaks to selection on pre-trends rather than selection on levels, and our design accommodates the latter through the district-by-grade fixed effects.

6 Results

Chain tutoring centers have a strong, positive, and persistent impact on math achievement and a more moderate impact on reading achievement. Following the opening of a first chain center, overall math (reading) test scores increase by 0.029 SDs (0.016 SDs). Figure 5a plots the event study coefficients from equation 3 separately for math and reading test scores. Both series exhibit flat pre-trends, lending credence to the parallel trends assumption necessary for a casual interpretation of the two-stage difference-in-difference design. In the first year following chain entry, test scores immediately rise and continue to increase at least five years out.

The observation that private tutoring has a larger impact on math scores than on reading scores is consistent with the curriculum focus of most tutoring centers. Evidence suggests that 60 to 75 percent of Kumon instruction is dedicated to mathematics. Ignoring potential differences in pedagogical quality or within-school spillovers across subjects, this distribution of instructional time should theoretically yield a roughly twice greater impact on math than on reading achievement. Our findings are consistent with this expectation.

A major identification concern is that tutoring centers may open in response to changes in local demand for tutoring services that are themselves correlated with future changes in student test scores (e.g., demographic shifts, changing school quality, increasing parental investment). To address these concerns, Figure 5b presents results that additionally control for grade-level enrollment, racial composition of the test-taking student body, and district-level unemployment and median income. Reassuringly, inclusion of these additional controls does not significantly change the magnitudes of our estimates. With controls, we estimate the impact of a first-chain center on math (reading) scores as 0.034 SDs (0.021 SDs), a slight increase from the specification without additional controls. Thus our measured impact does not appear to be driven by changing demographic or economic characteristics of the school district at the time of tutoring entry.

6.1 Heterogeneity by Race

The overall positive effect of chain tutoring centers masks large racial heterogeneity. Consistent with the patterns in tutoring consumption documented above, we find larger effects for White and Asian students than Black, Hispanic, and Native American students. Focusing on math test scores,we estimate that tutoring center entry drives large and persistent impacts on White and Asian students' test scores of a highly statistically significant 0.040 SD (Figure 6a). In contrast, the estimated impact for Black, Hispanic, and Native American students is less than half of that magnitude (0.018 SDs) and is not statistically distinguishable from 0. These figures also exhibit flat pretrends.[12]

Consistent with racial heterogeneity in effects across districts, tutoring centers also appear to widen within-district math test score gaps between White students and their Black and Hispanic counterparts. Figure 6b shows chain tutoring entry in a school district increases the gap between White and Black students' average math scores by a highly significant 0.037 SDs and the gap between White and Hispanic students' math scores by a marginally significant 0.020 SDs. While the estimates in Figure 6a are estimated on different subsamples of school districts (because SEDA suppresses subgroup test scores when there are fewer than 20 observations), the racial test score gap estimates presented in Figure 6b isolate the impact on students of different races in the same district. Thus the racial heterogeneity in tutoring chain impacts is driven by within-district effects, rather than compositional differences between our estimating samples or heterogeneity of tutoring center effects across locations that serve different student populations.

One potential mechanism for the large effects of private tutoring centers is within-school spillovers (Yao et al., 2024). These spillovers could arise if tutored students require less instructional attention, thereby freeing teachers to spend more time with other students, or if tutored students affect their classmates through peer learning. The spillovers channel has a distinctive implication: because private tutoring is disproportionately consumed by White and Asian students, tutoring centers should increase racial test score gaps more in settings where White and Asian students primarily interact with same-race peers, concentrating the spillovers by race.

We find evidence consistent with a spillover hypothesis. As seen in Figure 7a, in districts with above-median White-Black segregation, where White students are more likely to attend schools with other White students, the entry of a first tutoring chain increases the White-Black test score gap by 0.047 SDs. In contrast, the corresponding effect is less than half as large, 0.021 SDs in districts with below-median segregation, where Black students have greater exposure to White peers and

[12] Appendix Figure A.3a shows complementary results on reading test scores, which do not exhibit heterogeneity by race.

therefore more scope to benefit from any within-school spillovers. Figure 7b shows a similar pattern for the White-Hispanic gap. Tutoring-chain entry increases the gap by 0.028 standard deviations in highly segregated districts, but has no discernible effect in less segregated districts. These patterns suggest that the racial composition of students' peer environments mediates the effect of tutoring centers on racial achievement gaps, lending support to the hypothesis that there are spillovers from center-based private tutoring.

6.2 Heterogeneity by Other Student Characteristics and Estimates by Chain Type

In contrast to observed racial heterogeneity in effects, we do not see large disparities in the impact of tutoring by economic status. Figure 8a plots the event study coefficients of tutoring chain entry on math scores for non-economically disadvantaged and economically disadvantaged students. The two series closely track each other, exhibiting flat pre-trends and a modest rise in math test scores (0.02 SD) following chain-entry, and the within-district math achievement gap between non-economically disadvantaged and economically disadvantaged students exhibits no clear change following chain-entry (Figure 8b). These results are arguably puzzling given the heterogeneity in tutoring expenditure patterns by income described in section 4. Given tutoring attendance patterns, we expect non-economically disadvantaged students to benefit more from the introduction of chain tutoring centers. One potential explanation is that the measure of economic disadvantage captures over 40 percent of students, so that we are roughly comparing above and below median income students, rather than focusing on the top decile families where tutoring takeup is highest. As we will explore below, this null impact on economic heterogeneity is also somewhat sensitive to modeling choices. Under other reasonable constructions, we see non-economically disadvantaged students gaining more after tutoring entry, partially resolving this puzzle.

We additionally compute estimates of chain tutoring impacts for each racial category separately, for male and female students, and for elementary and middle grades. We present the average treatment effects for each subgroup in Column 1 of Table 3. Consistent with usage patterns, the impact of chain tutoring centers is larger for Asian students (0.064 SDs) than White students (0.030 SDs), but estimates for Asian students have larger standard errors due to the small number of districts with sufficient numbers of Asian students to report scores for that subgroup. Impacts are similar for Male (0.026 SDs) and Female (0.028 SDs) students. Younger students (those in grades 3-5) see slightly larger benefits from tutoring center entry than older students (grades 6-8). Column (2) of Table 3 replicates these results with additional controls for grade-level enrollment, grade-level racial shares, and district-level unemployment and median income. The results are largely unchanged, confirming that our estimates are not driven by changing district characteristics. Together, this analysis shows that tutoring chain entry creates differential effects by race, but less so by economic status, gender, or age.

We replicate this analysis for reading achievement, finding more modest impacts of tutoring centers on test scores. As seen in Table 4, the point estimate from our simplest specification is a statistically significant 0.016 standard deviations. Adding controls increases this to 0.021 standard deviations. Both point estimates for reading scores are roughly half the magnitude of their math counterparts. we do not find clear heterogeneity in the impacts on reading achievement by race or by economic status.[13] Point estimates for female students are consistently larger than those for male students. We cannot distinguish whether this is driven by gender differences in efficacy or takeup rates of tutoring in reading.

In contrast to our findings on chain tutoring centers, we find no impacts on math and reading scores from non-chain tutoring centers (those operating less than 10 locations) or from in-home chains that send tutors into students' homes.[14] The lack of detectable impacts from non-chain centers may reflect the role of standardized curricula and franchise oversight, features specific to chain tutoring centers, in delivering achievement gains. Tutoring center location may also be a poor-proxy for exposure to in-home tutoring chains, as employees traveling to a client's home need not be based near the registered business.

6.3 Extension to Continuous Measure of Tutoring Density

The empirical strategy above relies on measuring test scores in school districts that gained their first chain tutoring center during our sample frame. Consequently, our analysis is focused on the impacts of tutoring centers in a set of mid-sized, primarily suburban school districts. In this section, we show that our headline results hold using a continuous treatment measure of the number of chain

[13] Appendix figures A.3 - A.5 show event studies of the impact of tutoring center entry on reading scores by race and economic status.

[14] See Table A.1 for details.

tutoring centers per 1,000 children, a measure we extend to all school districts.[15]

Our central estimate of the effect of chain tutoring on math achievement does not depend on the choice of estimator or on the restricted estimation sample. We move from our two-stage DiD estimates presented in Column (2) of Table 3 to the all-district estimates in Column (5) by changing one feature of the specification at a time, holding the additional controls fixed throughout. Column (3) changes only the estimator, replacing two-stage difference-in-differences with two-way fixed effects. Because both use relative years to first entry, their coefficients are on the same scale, and the effect for all students moves from 0.034 to 0.040 standard deviations.

Column (4) changes only the treatment variable, replacing the relative years to first entry with a continuous, lagged count of chain centers per 1,000 children ages 5-17 in the school district. Here we report the coefficient on this treatment variable from a two-way fixed effects model. This alters the scale of the coefficient, which now measures the effect of one additional center per 1,000 children rather than the effect of first-entry, so the 0.044 SD estimate is not directly comparable to Columns (1) through (3). Column (5) changes only the sample, extending the continuous specification to all school districts, and returns an effect of 0.026 standard deviations. The smaller impact of chain tutoring centers in Column (5) relative to Column (4) may reflect diminishing returns to additional tutoring centers in large districts that contain many tutoring centers. The continuous, all-district design identifies the effect from within-district changes in tutoring density over time relative to other districts and years, rather than from the timing of a district's first center. The exogeneity of all such changes may be less plausible than our headline strategy. We report it because it is the only specification that uses the full set of districts, and the positive and statistically significant effect on math scores survives the extension.

The subgroup patterns documented above largely persist across these specifications: effects are largest for Asian and White students, similar for male and female students, and larger in grades 3-5 than in grades 6-8. The economic-status gap remains statistically indistinguishable from zero across every specification. The racial test score gap findings are somewhat sensitive to this progression. In the binary specifications estimated on the first-/never-chain sample, chain entry widens the within-district White-Black gap by 0.031 to 0.037 standard deviations and the White-Hispanic gap

[15] We use 1,000 children ages 5-17 from the ACS rather than 1,000 children enrolled in the school district, as the former captures students enrolled in both public and private schools.

by roughly 0.020 standard deviations. These gap effects get noiser once we adopt the continuous treatment measure, and the Black-White gap effect weakens further when extending the sample to all districts. Gap-widening effects may be specific to a community gaining its first center and diluted by the marginal centers opening in large, already tutoring-saturated urban districts of the full sample. This is consistent with the diminishing marginal returns evident in the overall effect.

6.4 Robustness

We perform three additional sets of robustness checks. First, we show that our results are not sensitive to alternate weighting schemes or the inclusion of state-by-year fixed effects.[16] Our baseline specification treats all district-grade-year observations with equal weight. Districts can vary substantially by size, so that unweighted regressions may not be representative of the underlying student population. Weighting by the number of student test takers in each district-grade-year has little impact on our overall estimates or estimates by race. We might also worry that tutoring chains tend to cluster center openings in places and times based on perceived demand, which could confound the impact of center openings with other geographically concentrated factors, such as economic or policy factors changing at the state level. Inclusion of state-by-year fixed effects, which control for such factors, again leaves our central estimates relatively unchanged. Models with both student weights and state-by-year fixed effects look similar to our baseline estimates, though with clearer heterogeneity in math impacts by student economic status.

Second, we show that our estimates of the overall impact of tutoring centers and racial heterogeneity are insensitive to the requirement that we observe test scores in a district-grade for at least ten of the eleven years from 2009 to 2019.[17] As we change alternative minimum-observation requirements from eleven years to two years of observed test-score data, the pooled estimate and racial heterogeneity patterns are stable across specifications even as the sample nearly doubles. The one margin along which the estimates shift is economic status. As the requirement is relaxed, the effect for non-economically-disadvantaged students holds between 0.021 SDs and 0.026 SDs, while the effect for economically-disadvantaged students declines from 0.025 to roughly 0.015, losing statistical significance and widening the gap by economic status from near zero under the most

[16] See Table A.2 for details.

[17] See Table A.3 for details.

restrictive requirement to nearly 0.01 SD.

Finally, our main estimates are similarly insensitive to an alternative geographic definition of treatment, which defines a school district as having a tutoring center if a tutoring chain operates within a 10-mile radius of the district centroid rather than within district boundaries.[18] Under this definition, we construct measures of tutoring chain entry and tutoring chain density per 1,000 children ages 5-17 and re-estimate the effect of chain tutoring centers on math and reading achievement. Two-stage difference-in-difference estimates under the 10-mile definition closely track those from the within-district specification, with larger overall impacts for math than reading and large racial heterogeneity in math impacts. The patterns by economic status are again more sensitive to our modeling decisions: in contrast to the within-district finding of comparable point treatment effects on math scores for non-economically and economically disadvantaged students, the 10-mile specification suggests gains are concentrated among non-disadvantaged students (0.037 vs. -0.002 standard deviations), in line with consumption patterns. Similarly, the continuous measure of treatment intensity suggests that additional tutoring centers have substantially larger impacts on the math scores of non-economically disadvantaged students. We interpret these results as qualifying our earlier finding that chain entry affects higher- and lower-income students similarly, highlighting the sensitivity of that finding to our empirical specification.

7 Treatment Effect Magnitudes and Cost Effectiveness

What do our district-level estimates imply about the effect of tutoring on the students who actually use these centers? Our event-study coefficients are intent-to-treat (ITT) effects, measuring the change in average test scores across all students in a treated district, the great majority of whom never enroll in a center. Assuming no spillovers and no lagged effects of prior tutoring, we recover a treatment-on-treated (TOT) effect by scaling the ITT by the share of students whose tutoring is induced by chain entry. This requires an estimate of take-up. We pursue two approaches: one based on household expenditure in the CEX, and one based on the physical capacity of the centers themselves.

We first infer take-up from the CEX. The opening of a district's first chain raises the density of

[18] See Table A.4 for details.

centers by 0.271 per 1,000 children (Figure 4). Because our outcomes cover grades 3–8—children roughly ages 8 to 13—we use the relationship between chain density and tutoring spending among households with children in that age range (Panel D of Figure 3), where each additional center per 1,000 children raises the share of households spending on tutoring by 31.1 percentage points. Multiplying these implies that first-chain entry raises take-up by about 8.4 percentage points, and dividing our ITT of 0.029 standard deviations (0.034 with controls) by this take-up yields a TOT effect of about 0.34 standard deviations (0.40 with controls).

This magnitude sits essentially at the central estimate from meta-analyses of the tutoring literature, where the average program raises achievement by about 0.4 standard deviations (Nickow et al., 2024). The take-up rate underlying it is, however, somewhat implausible. An 8.4 percentage point increase in tutoring from the arrival of a single center exceeds even the 5.0 percent tutoring rate among top-decile households nationally, and far exceeds what a center serving at most 100 to 130 students could physically deliver in a district of several thousand children. The likely explanation is that the CEX density-spending relationship is not causal: chains tend to locate where tutoring demand, including demand met through other channels, is already high, so the slope is inflated by unobserved covariates that our cross-sectional regression cannot absorb. If so, the CEX overstates the take-up caused by entry, and the 0.34 figure is best read as a lower bound on the effect for tutored students.

To avoid relying on a causal interpretation of the CEX, we measure take-up directly from center capacity. If the typical center serves 100 (from franchise disclosure documents) to 130 students (from Kumon's press release), then one additional center per 1,000 children mechanically tutors 10 to 13 percent of them, a take-up slope roughly a third of the 31 percentage points the CEX assigns, and one that depends only on enrollment counts rather than on expenditure patterns. Under this measure, first-chain entry raises take-up by 2.7 to 3.5 percentage points, and the implied TOT effect is correspondingly larger, on the order of 0.8 to 1.1 standard deviations. This figure is high relative to the experimental literature, and we suspect the true effect on tutored students is somewhat lower: to the extent that gains persist after a student stops attending or spill over to non-tutored classmates, the district-level effect is spread across more students than are tutored at any one time, so dividing by current take-up overstates the per-student effect. We therefore read the two approaches as bracketing the TOT effect, the CEX calculation providing a lower bound near 0.3 standard deviations and the capacity-based calculation an upper bound around 1.0.

The magnitudes of tutoring center effects also suggest that entry of such centers explains a substantial portion of the evolution of test scores in treated districts. From 2009-19 in such districts, overall math (reading) scores increased by 0.046 (0.030) standard deviations, so that our estimated center-entry effect of 0.029 (0.016) standard deviations explains 63 (53) percent of districts' overall improvement. For White and Asian students in those districts, math (reading) scores increased by 0.084 (0.053) standard deviations, so that our estimated center-entry effect of 0.040 (0.019) standard deviations explains 48 (36) percent of such students' improvement. Similar calculations suggest that center entry explains 44 (27) percent of the increased White-Black math (reading) gap in treated districts, and 95 (-26) percent of the change in the White-Hispanic math (reading) gap.[19] Such private investments outside of public school systems may have non-trivial effects on the outcomes by which school districts are evaluated by families and state accountability systems.

Tutoring centers appear strikingly cost-effective. Under our capacity-based estimate, a first chain induces 2.7 to 3.5 percent of students to take up tutoring, each spending about $290 per month based on CEX figures, or roughly $2,600 over a nine-month academic year. Spread across every pupil in the district, this amounts to roughly $70 to $90 per pupil. By comparison, Jackson and Mackevicius (2024) estimate that a $1,000 per-pupil increase in public school spending sustained over four years raises average test scores by 0.032 standard deviations. Our first-chain ITT of 0.029 standard deviations is nearly identical in magnitude, yet the tutoring expenditure that produces it is an order of magnitude smaller per pupil. Private tutoring centers thus purchase a district-average achievement gain at a small fraction of the public cost required to produce the same effect, albeit with the expense borne privately by families rather than by school systems.

This apparent efficiency, however, comes with a sharp distributional asymmetry. The gains accrue largely to students who take up tutoring, so the same spending that looks efficient in the aggregate widens racial achievement gaps rather than narrowing them. In contrast, Jackson and Mackevicius (2024) find the effects of public school spending are largest for economically disadvantaged students. Tutoring centers may be cost-effective in part because they target students with fewer challenges and more family support to complement the tutoring itself.

[19] From 2009-19, the White-Hispanic reading test score gap decreased by 0.065 standard deviations in treated districts, meaning that Hispanic scores improved relative to White scores.

8 Conclusion

This paper provides the first national, causal evidence on the effects of private tutoring chains on student achievement. Using two decades of business registry data linked to household spending and school district test scores, we show that tutoring usage is concentrated among Asian, White and high-income households, and that the opening of a district's first chain tutoring center raises district-wide math and reading scores. Math effects are driven largely by White and Asian students, widening within-district racial achievement gaps, particularly in racially segregated districts, consistent with a within-school spillover mechanism. Non-chain centers and in-home tutoring chains show no detectable effects, suggesting the standardized curricula and franchise oversight specific to chains, rather than tutoring generally, may drive these gains.

The implied magnitude of the effect on treated students is very high, raising the question of whether public schools could improve their own instruction by learning from these chains' approaches. Kumon's worksheet-based mastery model and Mathnasium's individualized math pathways implement principles (frequent low-stakes practice, work calibrated to a student's actual level rather than grade level, rapid feedback) that the broader tutoring literature has already identified as effective (Nickow et al., 2024). Chains add to this execution, delivering franchise standardization, instructor accountability, and diagnostic placement at a scale and consistency that individual schools and classrooms may struggle to achieve.

A second, related question is what to make of the demand these centers satisfy. Parents pay for tutoring centers likely because they perceive schools as not providing enough individualized instruction or practice time. That millions of families are willing to pay several hundred dollars a month for this suggests unmet demand for a service that public schools are positioned, in principle, to provide directly. Whether the binding constraint is school funding, instructional time, teacher capacity to differentiate within a classroom, or something else entirely is beyond what our design can answer, but the willingness-to-pay revealed by chain tutoring's growth is itself informative, highlighting a gap between what schools offer and what at least some parents believe their children need. This may be particularly relevant in the post-pandemic world in which public schools are increasingly struggling to attract high income and Asian and White families (Francis and Goodman, 2025).

Because usage and benefits are concentrated among already advantaged families, the same mechanism that makes tutoring centers valuable to their users currently makes them a force for widening, not narrowing, achievement gaps. Efforts to bring tutoring-style instruction into public schools at scale might ensure that the benefits of such pedagogical approaches are not limited to the families who can afford them in the private market.

References

Angrist, J. D., P. A. Pathak, and C. R. Walters (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics 5*(4), 1–27.

Aucejo, E. M. and T. F. Romano (2016). Assessing the effect of school days and absences on test score performance. *Economics of Education Review 55*, 70–87.

Bhatt, M. P., J. Guryan, S. A. Khan, M. LaForest-Tucker, and B. Mishra (2024). Can technology facilitate scale? evidence from a randomized evaluation of high dosage tutoring. Technical report, National Bureau of Economic Research.

Blazar, D., M. Boudreaux, S. Klees, J. King Rice, M. Titus, and J. Zhao (2026). Disparities in childhood human capital investments in the United States. *Nature Communications 17*, 2746.

Carlana, M. and E. La Ferrara (2025). Apart but connected: Online tutoring, cognitive outcomes, and soft skills. *American Economic Review 115*(10), 3487–3513.

Chen, M. H. (2026). The returns to private tutoring access. Working paper, University of Maryland.

Doepke, M. and F. Zilibotti (2019). *Love, Money, and Parenting: How Economics Explains the Way We Raise Our Kids.* Princeton, NJ: Princeton University Press.

Escueta, M., A. J. Nickow, P. Oreopoulos, and V. Quan (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature 58*(4), 897–996.

Figlio, D., K. L. Holden, and U. Ozek (2018). Do students benefit from longer school days? regression discontinuity evidence from florida's additional hour of literacy instruction. *Economics of Education Review 67*, 171–183.

Francis, A. and J. Goodman (2025). School enrollment shifts five years after the pandemic. *Economics of Education Review 109*, 102734.

Fryer, Roland G., J. (2014). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *The Quarterly Journal of Economics 129*(3), 1355–1407.

Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments*, Volume 2, pp. 95–322. Elsevier.

Gardner, J., N. Thakral, L. T. Tô, and L. Yap (2024). Two-stage differences in differences.

Goodman, J. (2014). Flaking out: Student absences and snow days as disruptions of instructional time. Technical report, National Bureau of Economic Research.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277. Themed Issue: Treatment Effect 1.

Guryan, J., J. Ludwig, M. P. Bhatt, P. J. Cook, J. M. Davis, K. Dodge, G. Farkas, R. G. Fryer Jr, S. Mayer, H. Pollack, et al. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review 113*(3), 738–765.

Jackson, C. K. and C. L. Mackevicius (2024). What impacts can we expect from school spending policy? evidence from evaluations in the united states. *American Economic Journal: Applied Economics 16*(1), 412–46.

Kawarazaki, H., M. Mahmud, Y. Sawada, and M. Seki (2023). Haste makes no waste: Positive peer effects of classroom speed competition on learning. *Oxford Bulletin of Economics and Statistics 85*(4), 755–772.

Kim, E., J. Goodman, and M. R. West (2025). Kumon in: The recent, rapid rise of private tutoring centers. *Education Finance and Policy 20*(3), 473–493.

Kornrich, S. and F. Furstenberg (2013). Investing in children: Changes in parental spending on children, 1972–2007. *Demography 50*(1), 1–23.

Kraft, M. A., J. A. List, J. A. Livingston, and S. Sadoff (2022). Online tutoring by college volunteers: Experimental evidence from a pilot program. In *AEA Papers and Proceedings*, Volume 112, pp. 614–618. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Kraft, M. A. and S. Novicoff (2024). Time in school: A conceptual framework, synthesis of the causal research, and empirical exploration. *American Educational Research Journal 61*(4), 724–766.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? evidence from developed and developing countries. *The Economic Journal 125*(588), F397–F424.

Liu, J., M. Lee, and S. Gershenson (2021). The short-and long-run impacts of secondary school absences. *Journal of Public Economics 199*, 104441.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review 109*(4), 1426–1460.

Nickow, A., P. Oreopoulos, and V. Quan (2024). The promise of tutoring for prek–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal 61*(1), 74–107.

Oreopoulos, P., C. Gibbs, M. Jensen, and J. Price (2024). Teaching teachers to use computer assisted learning effectively: Experimental and quasi-experimental evidence. Technical report, National Bureau of Economic Research.

Ramey, G. and V. A. Ramey (2010). The rug rat race. *Brookings Papers on Economic Activity 41*(1), 129–199.

Reardon, S. F., A. D. Ho, B. R. Shear, E. M. Fahle, D. Kalogrides, and J. Saliba (2024). Stanford Education Data Archive (Version 5.0). Retrieved from https://purl.stanford.edu/cs829jn7849.

Rivkin, S. G. and J. C. Schiman (2015). Instruction time, classroom quality, and academic achievement. *The Economic Journal 125*(588), F425–F448.

Sawada, Y., M. Mahmud, M. Seki, and H. Kawarazaki (2024). Fighting the learning crisis in developing countries: A randomized experiment of self-learning at the right level. *Economic Development and Cultural Change 72*(4), 1893–1921.

Schneider, D., O. P. Hastings, and J. LaBriola (2018). Income inequality and class divides in parental investments. *American Sociological Review 83*(3), 475–507.

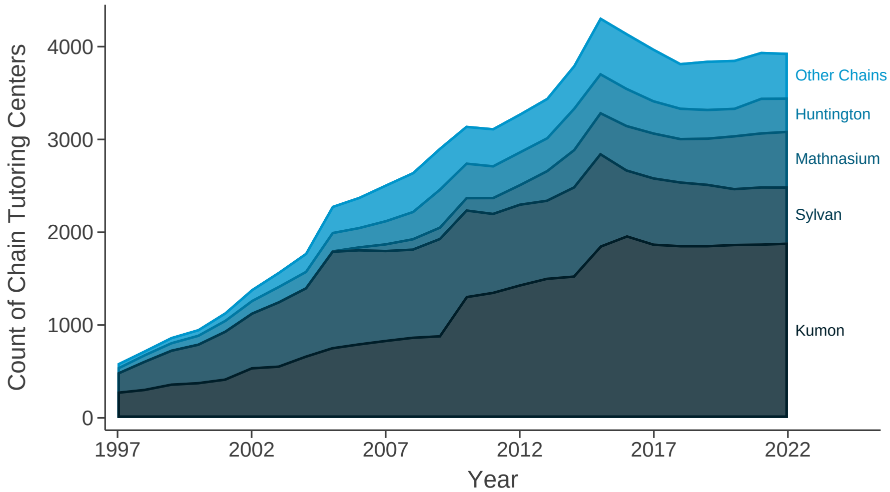
Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199. Themed Issue: Treatment Effect 1.

Thompson, P. N. (2021). Is four less than five? effects of four-day school weeks on student achievement in oregon. *Journal of Public Economics 193*, 104308.

U.S. Bureau of Labor Statistics (2024). Consumer Expenditure Survey, Public Use Microdata. U.S. Department of Labor. Accessed: 2025-11-19.

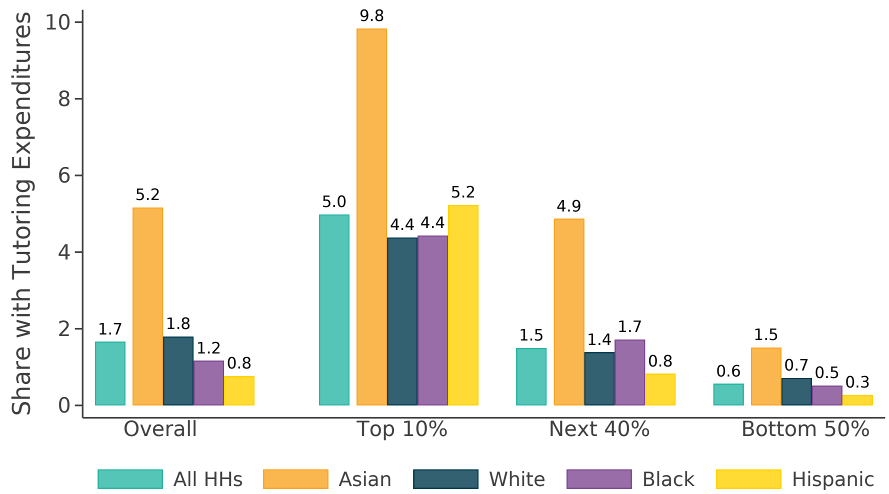
Yao, Z., H. Jiang, and J. Wang (2024). The spillover effect and external welfare of extracurricular tutoring: evidence from academic tutoring for middle-school students in china. *Applied Economics 56*(44), 5263–5288.

Figure 1: Count of Chain Tutoring Centers in the United States, 1997-2022



Notes: Figure 1 plots the total number of chain private tutoring centers in operation in the United States from 1997 to 2022. Series are shown separately for the four tutoring chains with the most locations: Kumon, Sylvan Learning, Mathnasium, and Huntington Learning, as well as for all other chains. Data are from Data Axle.

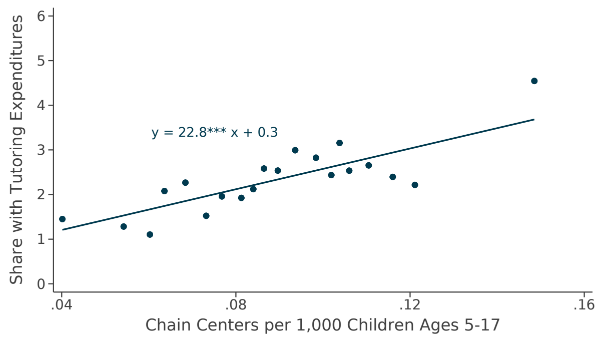
Figure 2: Share of Households with Tutoring Expenditure by Race/Ethnicity and Income Percentile



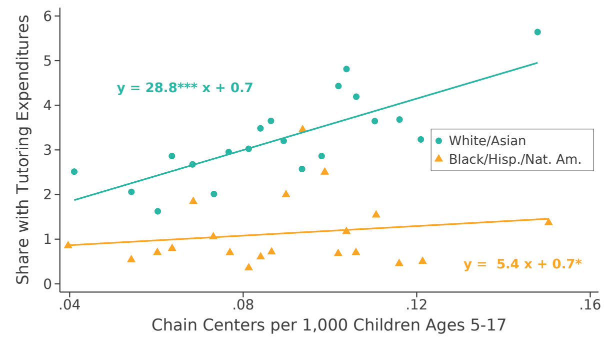
Notes: Figure 2 plots the share of households reporting non-zero expenditures on "Test preparation, tutoring services" in a given month, weighted by the CEX sample weights. Data come from the Consumer Expenditure Survey, 2009-2024. The leftmost group ('Overall') shows the share for all households. The three rightmost groups of bars compare households of different races within the same bin of the national income distribution. Top 10% refers to households with incomes above the 90th percentile on the national income distribution, Next 40% are households with incomes between the 50th and 90th income percentiles, and Bottom 50% are households with incomes below the 50th percentile. Within each income bin, shares are shown for all households ('All HHs') as well as separately by race/ethnicity. For more details on the sample construction, see section 3.2.

Figure 3: Tutoring Chain Center Density and Household Tutoring Participation

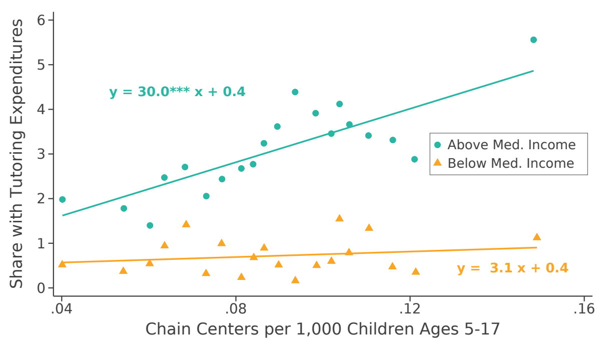
(a) All Households with Children Ages 6-17



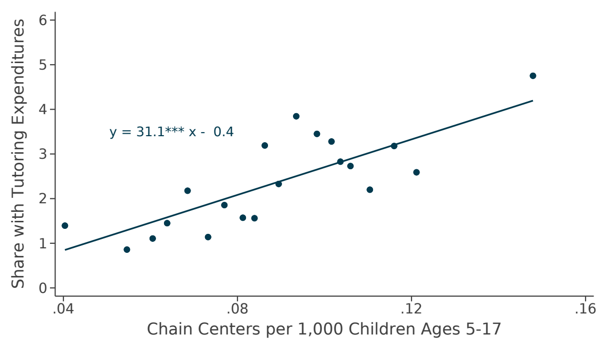
(b) By Race/Ethnicity



(c) By Income

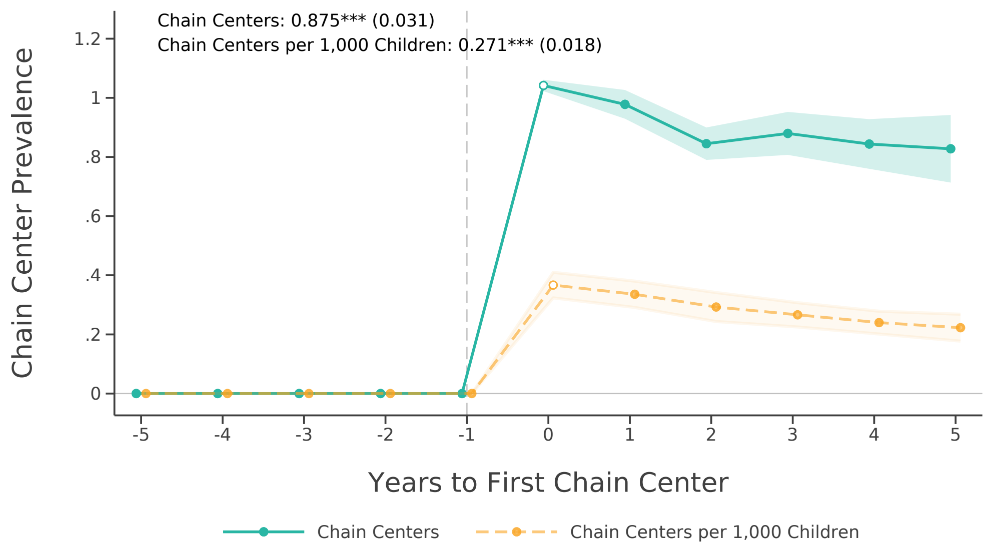


(d) All Households with Children Ages 8-13



Notes: Figure 3 plots the relationship between tutoring chain center prevalence and tutoring consumption across Metropolitan Statistical Areas (MSAs). For each MSA-year observation in the CEX from January 2009 to February 2024, tutoring chain density is calculated as the number of chains per 1,000 children ages 5–17. Household tutoring participation is an indicator equal to one if the household reports non-zero expenditures on "Test preparation or tutoring services" in a given month. Each panel reports the line of best fit and shows a binned scatter plot of household tutoring participation on MSA-year chain center density, weighted by CEX sample weights. Standard errors are clustered by household. Panel 3a shows the relationship for all households with children ages 6-17, our primary CEX sample. Panel 3b plots separate series White/Asian households and Black, Hispanic, and Native American households. Panel 3c plots separate series for households with income above and below the national median. Panel 3d restricts to households with children ages 8-13. The CEX sample excludes spending reported in June, July, and August. MSA-level geography is available for households in the 21 largest MSAs prior to 2016 and the 23 largest MSAs from 2016 onward. Tutoring chain center locations come from Data Axle and child population counts come from the American Community Survey (ACS). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

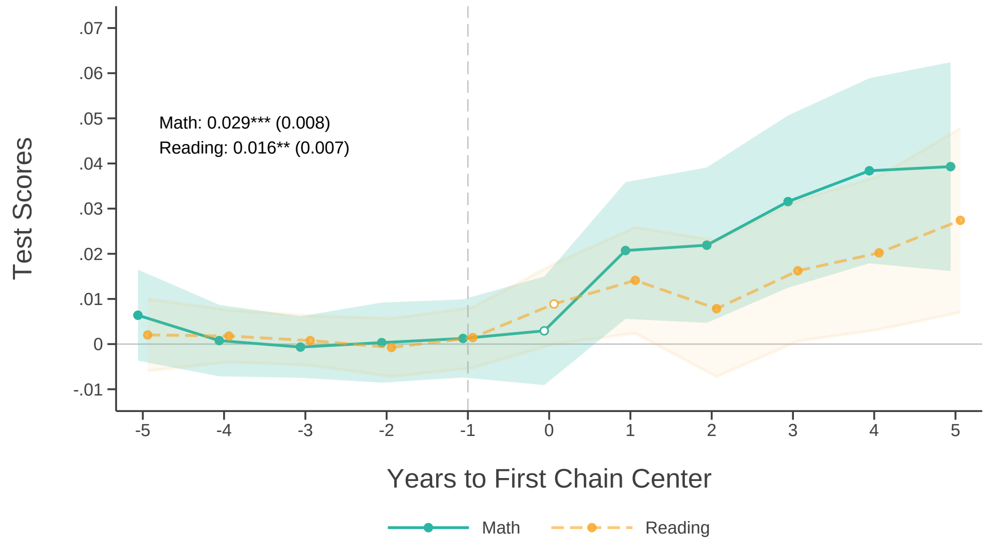
Figure 4: Impact of First Chain on Chain Center Prevalence



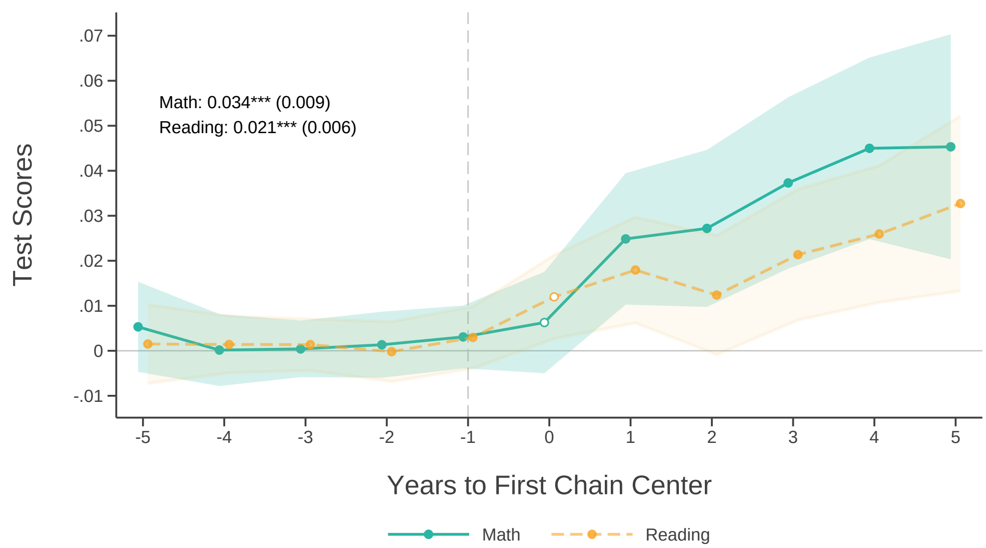
Notes: Figure 4 plots event-time estimates of the impact of first chain entry on the number of chain tutoring centers operating in a school district. The solid blue line shows event time coefficients on the number of chains, the dashed orange line shows event time coefficients for the number of chains per 1,000 children ages 5-17. Estimates include district and year fixed effects, but no additional controls. See equation 3 for the estimating equation. The average treatment effect is calculated using equation 4, omitting period 0. Standard errors are bootstrapped and clustered by school districts, the shaded region shows 95% confidence intervals. Tutoring chain locations come from Data Axle. School district boundaries come from SEDA. Number of children ages 5-17 comes from the ACS. The sample contains first-chain and never-chain school districts from 2009-2019 with at least 10 years of observed test score data. For more details on sample construction, see section 5.1. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 5: Effect of First Tutoring Chain on Test Scores, All Students

(a) Baseline



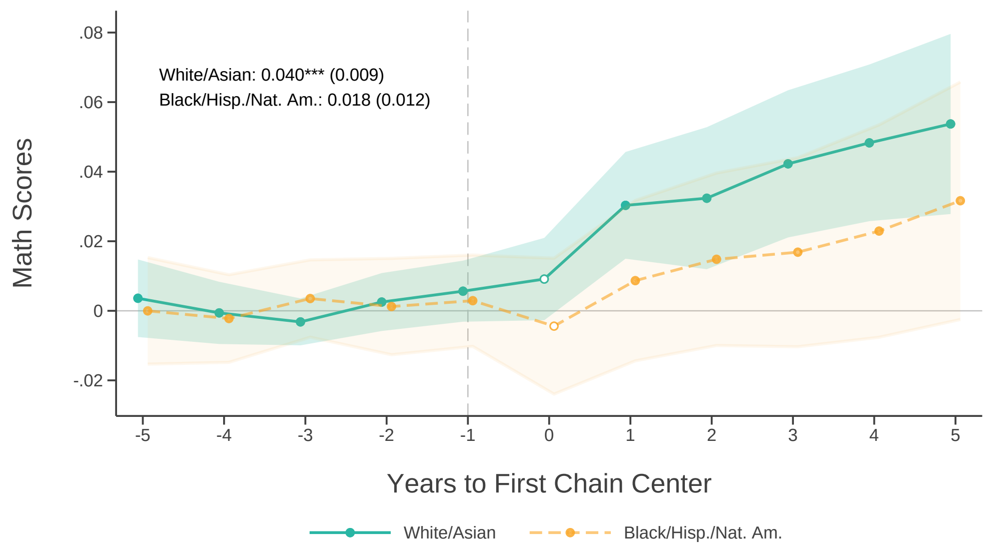
(b) With Additional Controls



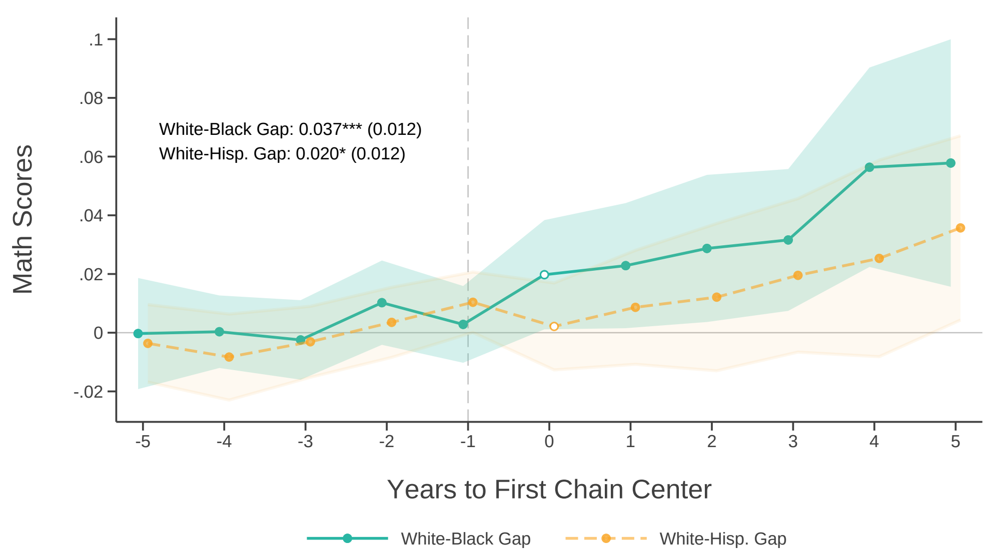
Notes: Figure 5 plots event-time estimates of chain tutoring entry on district-grade test scores. The solid blue line in panel 5a plots event-study coefficients from math test score regressions, the dashed orange line plots coefficients on reading test scores. Estimates include district-grade and year fixed effects, but no additional controls. See equation 3 for the estimating equation. The average treatment effects are calculated using equation 4, omitting period 0. Panel 5b replicates panel 5a but additionally controls for grade-level enrollment and racial shares, and district-level unemployment and median income. Standard errors are bootstrapped and clustered by school districts, the shaded region shows 95% confidence intervals. Tutoring chain locations come from Data Axle. School district boundaries, test scores, and additional controls come from SEDA. The sample contains first-chain and never-chain school districts from 2009-2019 with at least 10 years of observed test score data. For more details on sample construction, see section 5.1. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 6: Racial Heterogeneity in Math Effects

(a) By Race/Ethnicity



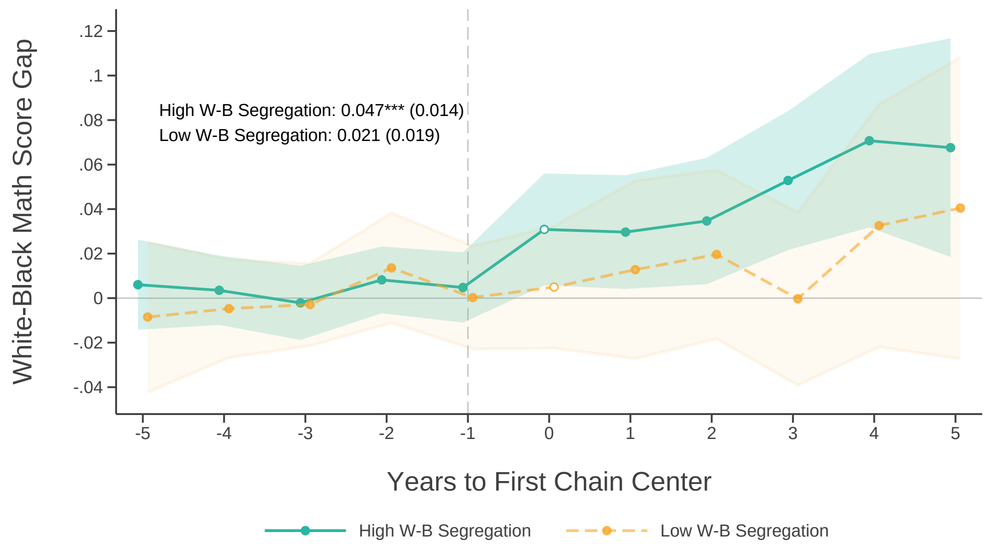
(b) Racial Test Score Gaps



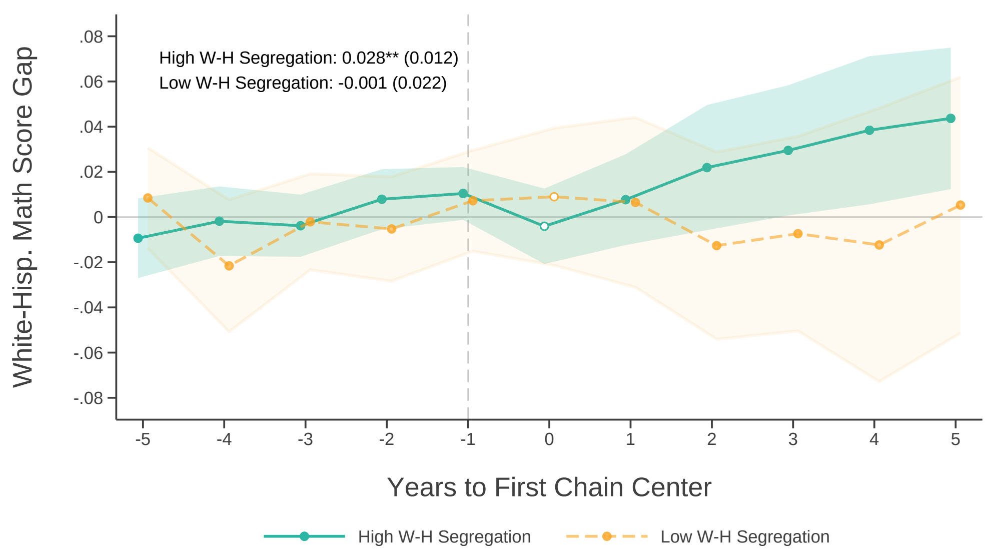
Notes: Figure 6 plots event-time estimates of chain tutoring entry on district-grade math test scores by race. In panel 6a, the solid blue line plots estimates using White and Asian students' test scores as the outcome. The dashed orange line is estimated with the test scores of Black, Hispanic, and Native American students. In panel 6b, the blue line plots estimates of the within-district gap between White and Black students' math test scores, the orange line plots estimates of the within-district gap between White and Hispanic students' math test scores. Estimates include district-grade and year fixed effects, but no additional controls. See equation 3 for the estimating equation. The average treatment effects are calculated using equation 4, omitting period 0. Standard errors are bootstrapped and clustered by school districts, the shaded region shows 95% confidence intervals. Tutoring chain locations come from Data Axle. School district boundaries and test scores come from SEDA. The sample contains first-chain and never-chain school districts from 2009-2019 with at least 10 years of observed test score data. For more details on sample construction, see section 5.1. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 7: Effect of First Tutoring Chain on Math Score Gaps by Racial Segregation

(a) White-Black Math Score Gap



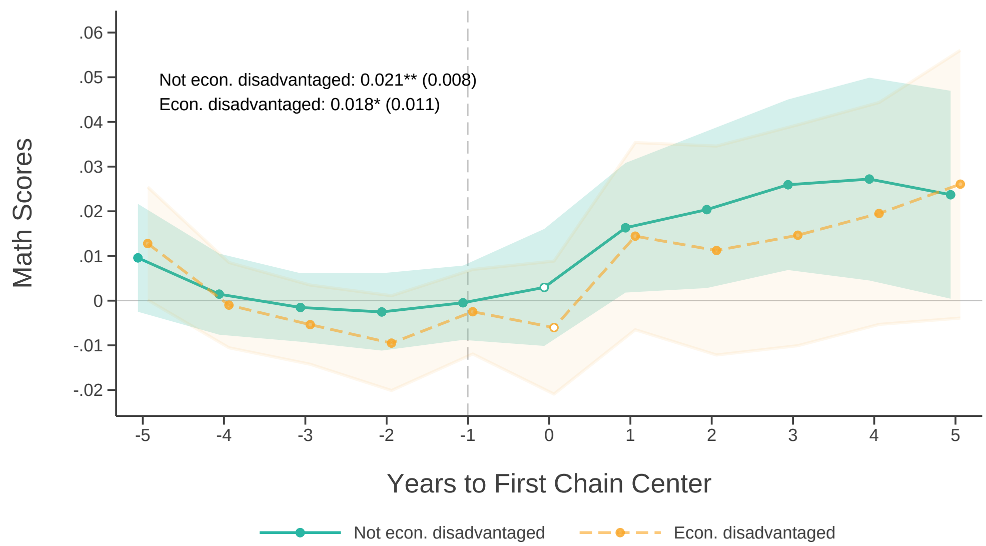
(b) White-Hispanic Math Score Gap



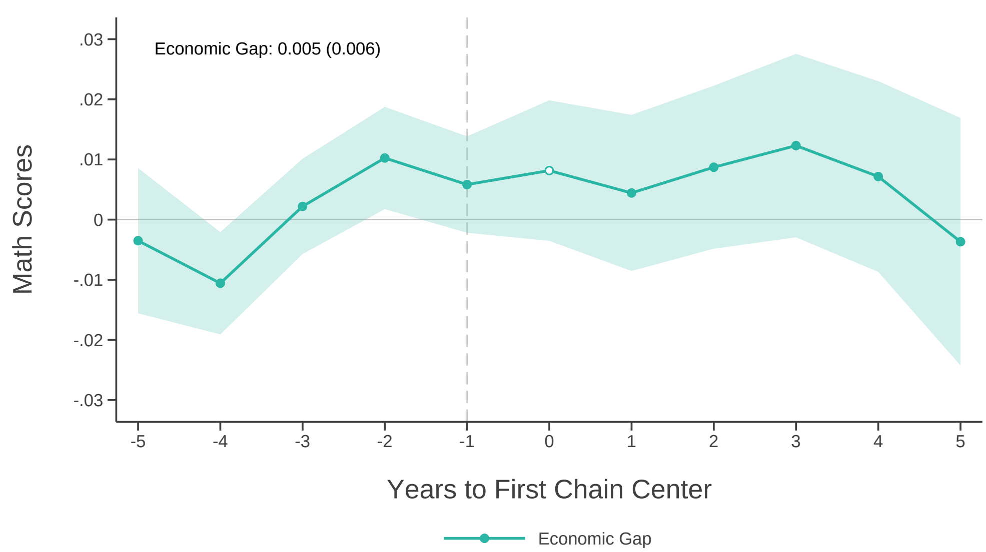
Notes: Figure 7 plots event-time estimates of the effect of a first chain tutoring center on within-district math test score gaps, separately by level of racial segregation. Panel 7a shows estimates for the White-Black math score gap; panel 7b shows estimates for the White-Hispanic gap. In each panel, district-grades are split at the median of the relevant information index based on the district-grade mean level of segregation across all observed years. The solid blue line includes districts with above-median segregation and the dashed orange line uses below-median districts. Estimates include district-grade and year fixed effects but no additional controls. See equation 3 for the estimating equation; average treatment effects are calculated using equation 4, omitting period 0. Standard errors are bootstrapped and clustered by school districts, the shaded region shows 95% confidence intervals. Tutoring chain locations come from Data Axle. School district boundaries and test scores come from SEDA. The sample contains first-chain and never-chain school districts from 2009–2019 with at least 10 years of observed test score data. For more details on sample construction, see section 5.1. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 8: Economic Heterogeneity of First Tutoring Chain on Math Test Scores

(a) By Economic Disadvantaged Status



(b) Economic Test Score Gap



Notes: Figure 8 plots event-time estimates of chain tutoring entry on district-grade math test scores by economic status. In panel 8a, the blue line uses the math test scores of non-economically disadvantaged students as the outcome, the orange line uses the math test scores of economically disadvantaged students. In panel 8b, the outcome is the within-district math score gap between economically disadvantaged and non-economically disadvantaged students. Estimates include district-grade and year fixed effects, but no additional controls. See equation 3 for the estimating equation. The average treatment effects are calculated using equation 4, omitting period 0. Standard errors are bootstrapped and clustered by school districts, the shaded region shows 95% confidence intervals. Tutoring chain locations come from Data Axle. School district boundaries and test scores come from SEDA. The sample contains first-chain and never-chain school districts from 2009-2019 with at least 10 years of observed test score data. For more details on sample construction, see section 5.1. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1: Counts of Chain Tutoring Brands

Brand Name	Subject (1)	Unique Locations from 2009–2019 (2)	Unique Locations in 2009 (3)	School Districts Treated (4)
Kumon	Math and reading	2,658	892	163.5
Sylvan Learning	Math and reading	1,780	1,048	139
Mathnasium	Math only	723	122	34.5
Huntington Learning	Math and reading	705	409	16
Eye Level Learning	Math and reading	271	0	8
Tutoring Club	Math and reading	215	160	10
The Tutoring Center	Math and reading	133	32	14
C2 Education	Math and reading	116	37	3
LearningRx	Math and reading	89	25	3
KnowledgePoints	Math and reading	63	55	1
The Learning Center	Math and reading	39	22	3
Enopi	Math and reading	38	9	1
JEI Learning Center	Math and reading	31	16	1
Lindamood-Bell	Reading only	31	20	1
Oxford Learning	Math and reading	23	22	0
Masonic Learning Center	Reading only	17	15	0
Reading Town	Reading only	17	7	0
Gideon Math & Reading	Math and reading	15	5	2
Math Monkey	Math only	14	12	0
Omega Learning	Math and reading	13	1	0
Aloha Mind Math	Math only	13	0	0
Grade Power Learning	Math and reading	9	0	2
Russian Math	Math only	8	1	0
Chain Totals		7,071	2,910	402

Notes: Table 1 presents the number and characteristics of chain tutoring center brands. Column (1) reports the primary subject(s) offered. Column (2) reports the cumulative number of unique center locations observed across the sample period from 2009 to 2019, and Column (3) reports the number of unique active locations in 2009. Column (4) reports the number of school districts treated by each chain brand. A school district is defined as treated when a chain opens a location within its boundaries for the first time. For more details on the construction of our estimation sample, see section 5.1. Fractional values in Column (4) (e.g., 163.5 for Kumon, 34.5 for Mathnasium) occur if multiple brands open in a school district in the year of first entry; the treatment count is proportionally split among those brands. The bottom row aggregates these metrics across all listed chains. Tutoring center location and operation data come from Data Axle; school district boundaries come from SEDA.

Table 2: Summary Statistics of School Districts in 2009

	All Districts (1)	First Chain (2)	Never Chain (3)
Enrollment and Achievement			
Number of Children Ages 5-17	4,867	5,831	2,236
Enrollment per Grade	344	432	163
Math Test Score (Std.)	0.007	0.179	-0.028
Reading Test Score (Std.)	-0.002	0.167	-0.036
Demographics			
Percent White	74.9%	71.0%	77.6%
Percent Hispanic	10.7%	13.2%	9.7%
Percent Black	9.8%	9.7%	8.7%
Percent Asian	2.1%	4.7%	1.2%
Percent Econ. Disadvantaged	44.1%	34.0%	46.1%
Share in Suburban Schools	23.3%	50.2%	16.0%
Employment and Income			
Median Income	52,917	65,907	50,066
Unemployment Rate	6.8%	6.2%	6.8%
Segregation			
White-Black Information Index	0.056	0.060	0.043
White-Hispanic Information Index	0.045	0.052	0.032
Economic Information Index	0.028	0.043	0.016
Unique Districts	8,943	402	6,959
Unique District-Grades	43,203	1,924	33,758

Notes: Table 2 presents summary statistics of our primary analysis sample in 2009, as described in section 5.1. Data cover grades 3-8 of all U.S. public school districts for which we observe test score data ten times between 2009 and 2019. Column (1) reports the mean characteristics for all district-grades. Column (2) reports mean characteristics for districts that gain their first chain tutoring center between 2010 and 2019, the treated sample. Column (3) reports mean characteristics for districts that never have a chain tutoring center between 1997 and 2019. Number of Children Ages 5-17 comes from the American Community Survey (ACS). All other variables come from SEDA. Demographic variables are based on the share of math test-takers in a district-grade. Variables in the "Employment and Income" panel are reported at the district level

Table 3: Tutoring Chain Centers on Math Test Scores by Subgroup, Model Comparison

Model	2-Stage DiD (1)	2-Stage DiD (2)	TWFE (3)	TWFE per 1,000 (4)	TWFE per 1,000 (5)
All Students	0.029***	0.034***	0.040***	0.044**	0.026***
	(0.008)	(0.009)	(0.009)	(0.017)	(0.009)
Race/Ethnicity					
White	0.030***	0.034***	0.040***	0.048***	0.018*
	(0.010)	(0.010)	(0.010)	(0.018)	(0.010)
Asian	0.064**	0.047*	0.114***	0.063	0.041*
	(0.030)	(0.026)	(0.041)	(0.051)	(0.023)
Black	0.003	0.009	0.025	0.068	0.006
	(0.016)	(0.016)	(0.019)	(0.059)	(0.024)
Hispanic	0.015	0.018	0.020	0.055	0.029
	(0.014)	(0.014)	(0.015)	(0.046)	(0.023)
Economic Status					
Non-Econ. Disadvantaged	0.021**	0.025***	0.029***	0.035**	0.019**
	(0.008)	(0.008)	(0.010)	(0.017)	(0.009)
Econ. Disadvantaged	0.018*	0.025**	0.024**	0.046	0.027*
	(0.011)	(0.010)	(0.011)	(0.032)	(0.014)
Gender					
Male	0.026***	0.030***	0.036***	0.051***	0.025**
	(0.009)	(0.010)	(0.010)	(0.019)	(0.010)
Female	0.028***	0.033***	0.038***	0.069***	0.034***
	(0.009)	(0.009)	(0.010)	(0.019)	(0.010)
Grade Level					
Grades 3-5	0.033***	0.038***	0.044***	0.048**	0.032***
	(0.010)	(0.010)	(0.011)	(0.020)	(0.011)
Grades 6-8	0.022**	0.028***	0.032***	0.038**	0.018**
	(0.010)	(0.009)	(0.010)	(0.018)	(0.009)
Test Score Gaps					
White-Black Gap	0.037***	0.035***	0.031***	0.027	-0.002
	(0.012)	(0.012)	(0.012)	(0.046)	(0.020)
White-Hispanic Gap	0.020*	0.019	0.022*	-0.021	-0.017
	(0.012)	(0.012)	(0.012)	(0.027)	(0.015)
White-Asian Gap	-0.022	-0.016	-0.038	-0.037	-0.044**
	(0.026)	(0.029)	(0.041)	(0.045)	(0.022)
Econ. Status Gap	0.005	0.004	0.009	0.022	0.003
	(0.006)	(0.006)	(0.007)	(0.017)	(0.010)
Number of Obs.	381,356	381,356	381,356	381,356	461,015
Add. Controls		×	×	×	×
Sample	First/Never Chain	First/Never Chain	First/Never Chain	First/Never Chain	All Districts

Notes: Table 3 presents estimates of the impact of tutoring chain centers on math test scores across models and student subgroups. Each cell is a separate regression. All estimates include calendar-year and district-grade fixed effects. Column (1) presents the average treatment effect from Equation 4, calculated using two-stage difference-in-differences, estimated on our sample of first-chain and never-chain districts from 2009-2019, and omitting period 0. Column (2) replicates Column (1) with additional controls for grade-level enrollment, grade-level racial shares, district-level unemployment and median income. Column (3) replicates Column (2), switching the estimation approach to standard Two Way Fixed Effects (TWFE). Column (4) defines treatment as a continuous measure of chain tutoring centers per 1,000 children ages 5-17, and reports the coefficient on this treatment variable from a TWFE model. Column (5) replicates Column (4), extending the analysis sample to all school districts from 2009-2019. Columns (1) and (2) use bootstrapped standard errors and all standard errors are clustered by school district. "Test Score Gaps" refer to the within-district gap in test scores between student subgroups. Test score data and additional controls come from SEDA. Tutoring center locations come from Data Axle. Counts of child population come from the ACS. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

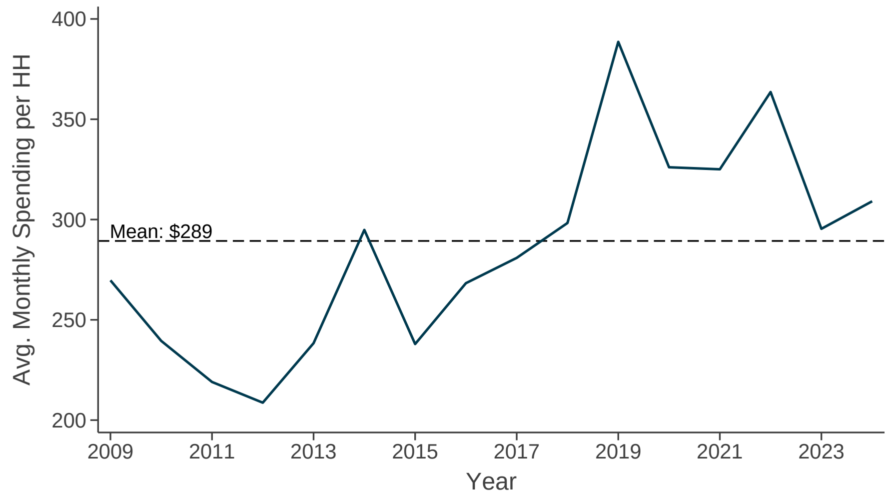
Table 4: Tutoring Chain Centers on Reading Test Scores by Subgroup, Model Comparison

Model	2-Stage DiD (1)	2-Stage DiD (2)	TWFE (3)	TWFE per 1,000 (4)	TWFE per 1,000 (5)
All Students	0.016**	0.021***	0.023***	0.026*	0.011
	(0.007)	(0.006)	(0.007)	(0.015)	(0.008)
Race/Ethnicity					
White	0.011	0.014*	0.013	0.026	0.002
	(0.007)	(0.007)	(0.008)	(0.016)	(0.009)
Asian	-0.021	-0.010	-0.006	-0.104**	0.007
	(0.021)	(0.023)	(0.028)	(0.048)	(0.020)
Black	0.006	0.009	0.015	0.032	-0.017
	(0.013)	(0.011)	(0.014)	(0.044)	(0.020)
Hispanic	0.008	0.012	0.008	-0.003	0.000
	(0.013)	(0.011)	(0.013)	(0.044)	(0.022)
Economic Status					
Non-Econ. Disadvantaged	0.010	0.013*	0.013	0.026*	0.010
	(0.008)	(0.007)	(0.008)	(0.015)	(0.008)
Econ. Disadvantaged	0.013	0.018**	0.015*	0.004	0.007
	(0.008)	(0.008)	(0.009)	(0.026)	(0.012)
Gender					
Male	0.012	0.016**	0.016**	0.017	0.003
	(0.007)	(0.007)	(0.008)	(0.017)	(0.008)
Female	0.020***	0.025***	0.028***	0.043**	0.022**
	(0.007)	(0.007)	(0.008)	(0.019)	(0.010)
Grade Level					
Grades 3-5	0.018**	0.022***	0.025***	0.024	0.016*
	(0.008)	(0.008)	(0.008)	(0.017)	(0.009)
Grades 6-8	0.014**	0.019***	0.020**	0.027	0.005
	(0.007)	(0.007)	(0.008)	(0.018)	(0.009)
Test Score Gaps					
White-Black Gap	0.016	0.016	0.011	-0.016	0.005
	(0.012)	(0.012)	(0.013)	(0.032)	(0.016)
White-Hispanic Gap	0.017	0.016	0.017	-0.026	-0.009
	(0.012)	(0.011)	(0.011)	(0.025)	(0.017)
White-Asian Gap	0.004	-0.011	0.013	0.045	-0.020
	(0.023)	(0.024)	(0.028)	(0.039)	(0.017)
Econ. Status Gap	0.001	-0.001	-0.000	0.017	0.012
	(0.006)	(0.006)	(0.006)	(0.017)	(0.009)
Number of Obs.	418,618	418,618	418,738	418,738	508,907
Add. Controls		×	×	×	×
Sample	First/Never Chain	First/Never Chain	First/Never Chain	First/Never Chain	All Districts

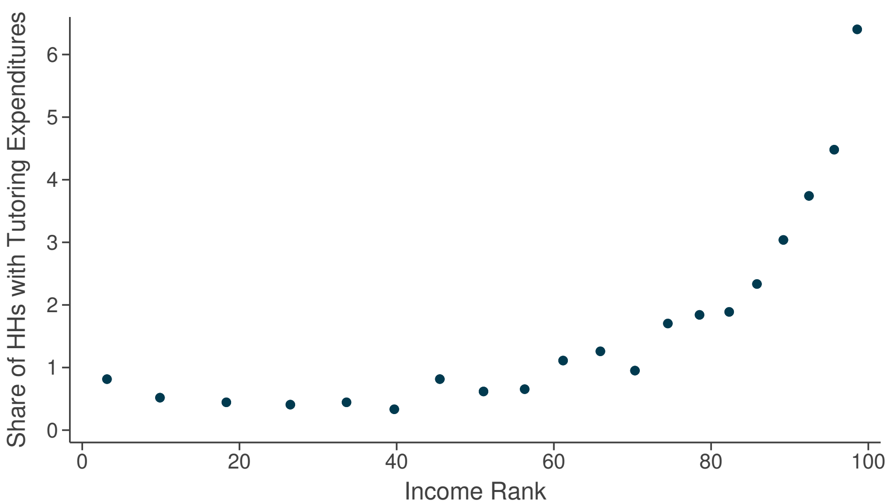
Notes: Table 4 presents estimates of the impact of tutoring chain centers on reading test scores across models and student subgroups. Each cell is a separate regression. All estimates include calendar-year and district-grade fixed effects. Column (1) presents the average treatment effect from Equation 4, calculated using two-stage difference-in-differences, estimated on our sample of first-chain and never-chain districts from 2009-2019, and omitting period 0. Column (2) replicates Column (1) with additional controls for grade-level enrollment, grade-level racial shares, district-level unemployment and median income. Column (3) replicates Column (2), switching the estimation approach to standard Two Way Fixed Effects (TWFE). Column (4) defines treatment as a continuous measure of chain tutoring centers per 1,000 children ages 5-17, and reports the coefficient on this treatment variable from a TWFE model. Column (5) replicates Column (4), extending the analysis sample to all school districts from 2009-2019. Columns (1) and (2) use bootstrapped standard errors and all standard errors are clustered by school district. "Test Score Gaps" refer to the within-district gap in test scores between student subgroups. Test score data and additional controls come from SEDA. Tutoring center locations come from Data Axle. Counts of child population come from the ACS. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.1: Tutoring Expenditure Patterns

(a) Monthly Tutoring Expenditure Among HHs with Tutoring Expenditures

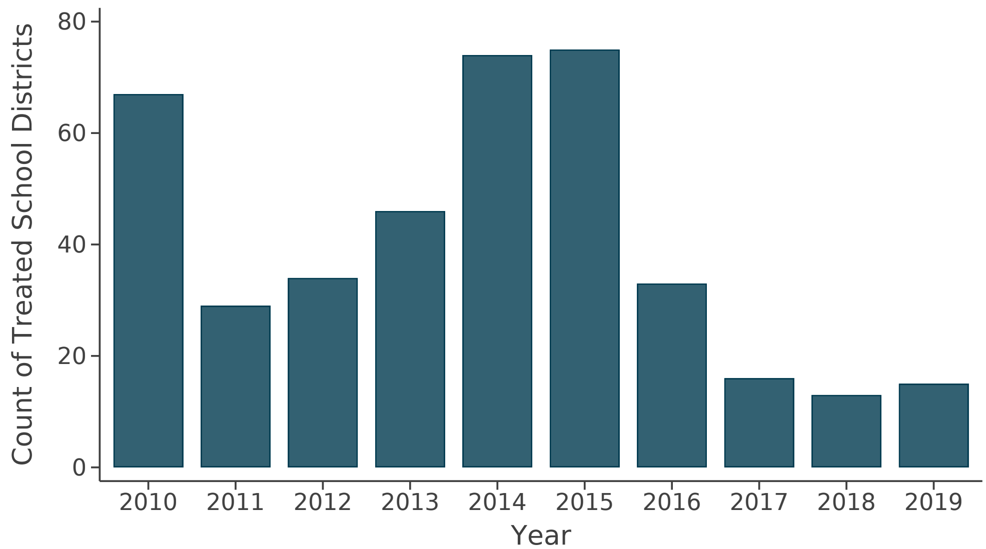


(b) Share of HHs with Tutoring Expenditure by Income Rank



Notes: Figure A.1a plots the average monthly spending of households on tutoring services by year among households with nonzero tutoring expenditures. The dashed line shows the overall mean, $289. Dollar values are inflation-adjusted to 2015 dollars. Figure A.1b plots the share of households reporting nonzero tutoring expenditures for each ventile of the national income distribution. All statistics use CEX sample weights. Data come from the Consumer Expenditure Survey, 2009-2024. For more details on the sample construction, see section 3.2.

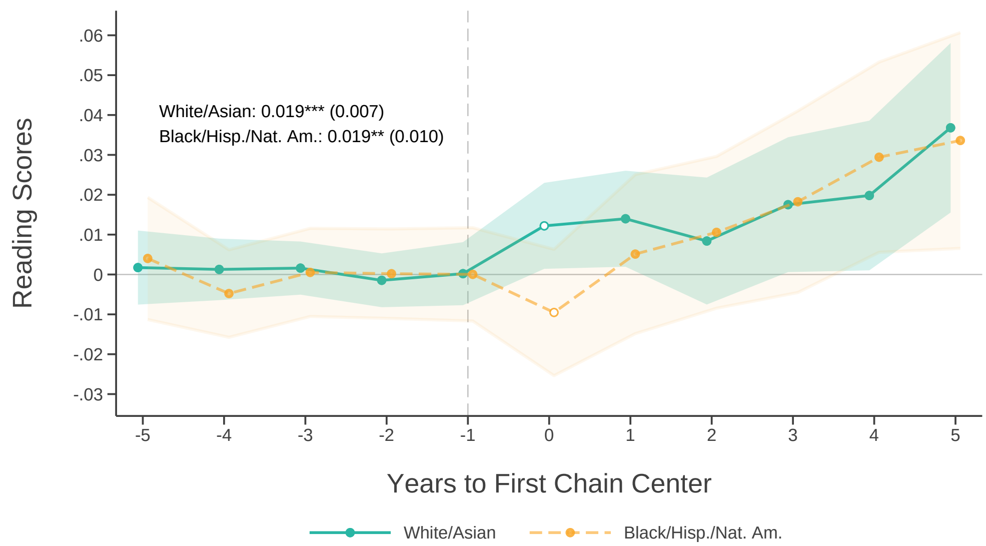
Figure A.2: Count of Treated Districts by Year



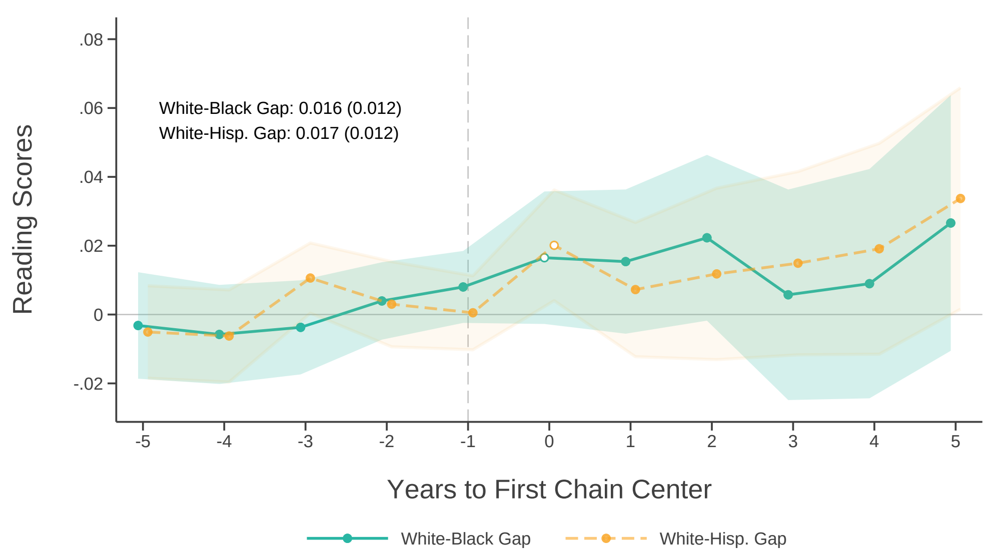
Notes: Figure A.2 plots the number of school districts that gained their first chain tutoring center by year from 2010 to 2019. Tutoring chain locations come from Data Axle. School district boundaries come from SEDA.

Figure A.3: Racial Heterogeneity of First Tutoring Chain on Reading Test Scores

(a) By Race/Ethnicity



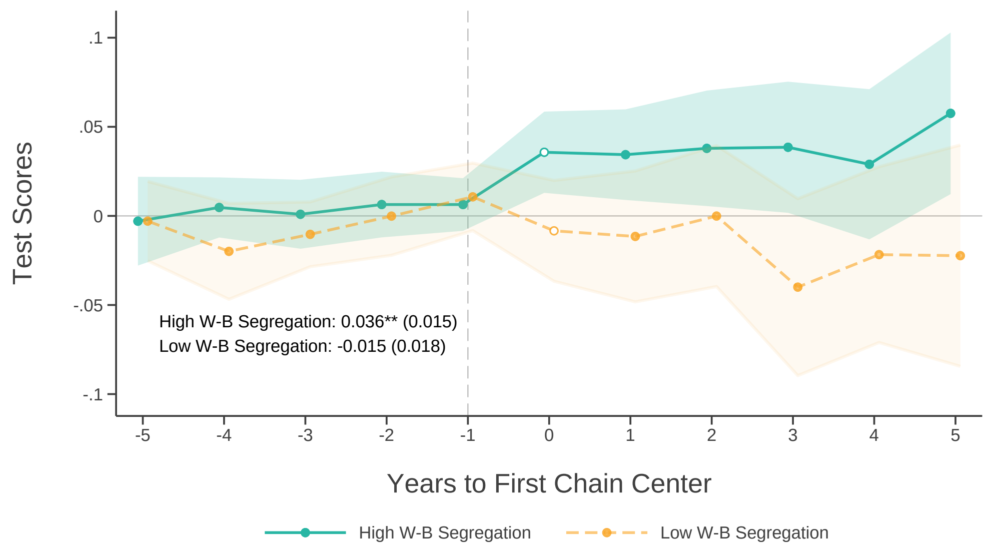
(b) Racial Test Score Gaps



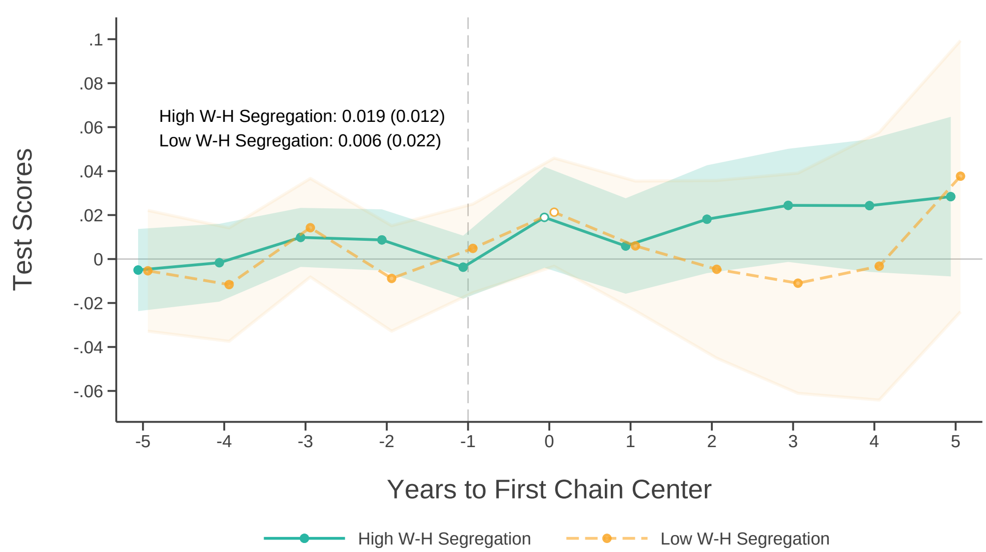
Notes: Figure A.3 replicates Figure 6 using reading test scores. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.4: Effect of First Tutoring Chain on Reading Score Gaps by Racial Segregation

(a) White-Black Reading Score Gap



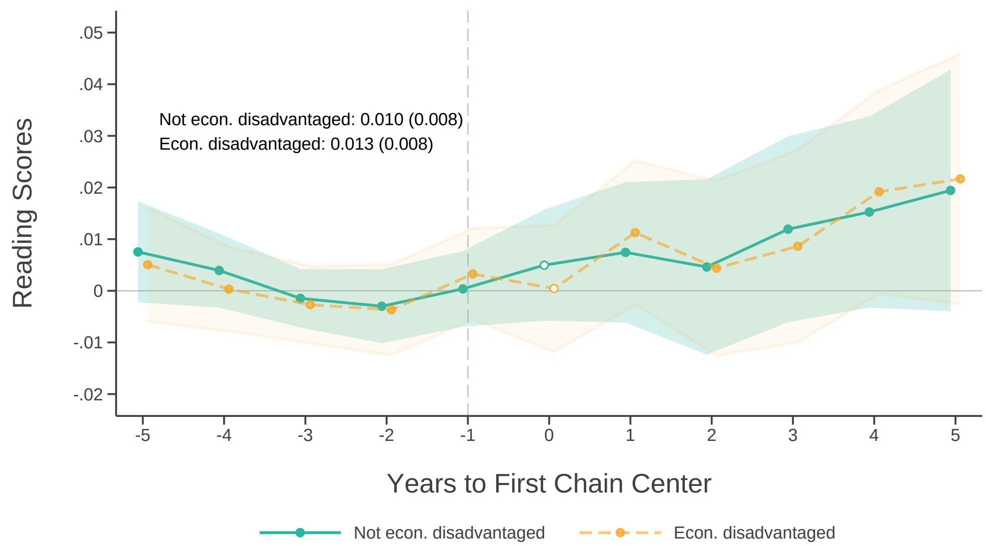
(b) White-Hispanic Reading Score Gap



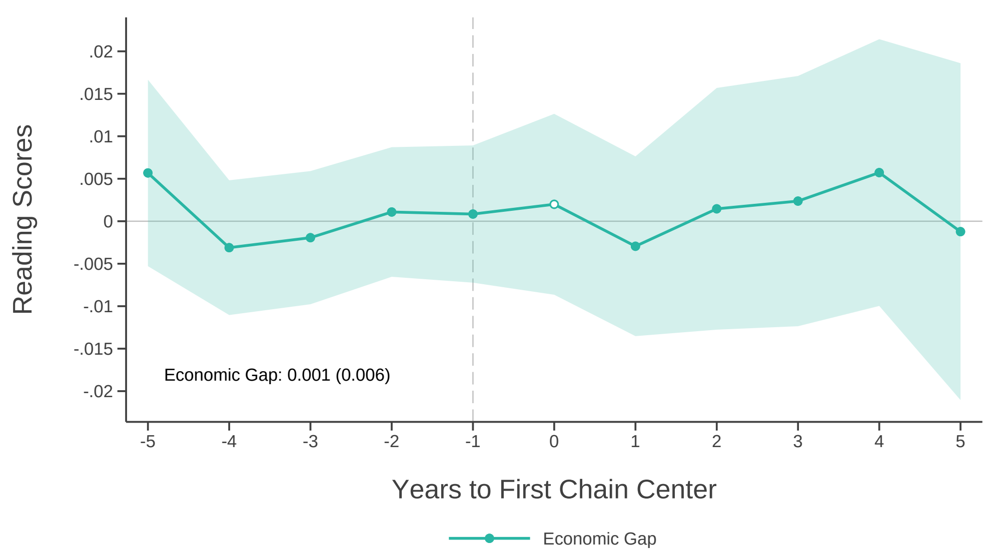
Notes: Figure A.4 replicates Figure 7 using reading test scores. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.5: Economic Heterogeneity of First Tutoring Chain on Reading Test Scores

(a) By Economic Disadvantaged Status



(b) Economic Test Score Gap



Notes: Figure A.5 replicates Figure 8 using reading test scores. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.1: Effect of Non-Chain and In-Home Tutoring Centers on Test Scores

Model	Non-Chains Math (1)	Non-Chains Reading (2)	In-Home Chains Math (3)	In-Home Chains Reading (4)
All Students	-0.002	0.003	0.009	0.005
	(0.009)	(0.006)	(0.011)	(0.008)
Race/Ethnicity				
White	-0.007	0.002	0.015	0.002
	(0.009)	(0.007)	(0.010)	(0.008)
Asian	-0.012	-0.044	0.026	-0.020
	(0.039)	(0.030)	(0.019)	(0.015)
Black	0.006	0.017	0.001	-0.005
	(0.017)	(0.014)	(0.019)	(0.013)
Hispanic	0.018	0.025**	0.010	0.008
	(0.015)	(0.012)	(0.012)	(0.010)
Economic Status				
Non-Econ. Disadvantaged	-0.013	-0.003	0.016	0.005
	(0.009)	(0.007)	(0.011)	(0.009)
Econ. Disadvantaged	-0.003	0.007	-0.003	0.009
	(0.010)	(0.007)	(0.011)	(0.009)
Gender				
Male	-0.008	-0.002	0.014	0.003
	(0.009)	(0.007)	(0.011)	(0.007)
Female	-0.008	0.007	0.005	0.005
	(0.009)	(0.006)	(0.010)	(0.008)
Grade Level				
Grades 3-5	-0.008	-0.001	0.008	0.000
	(0.010)	(0.009)	(0.011)	(0.010)
Grades 6-8	0.006	0.007	0.008	0.010
	(0.009)	(0.006)	(0.012)	(0.008)
Test Score Gaps				
White-Black Gap	-0.018*	-0.022**	0.024**	0.012
	(0.010)	(0.011)	(0.012)	(0.009)
White-Hispanic Gap	-0.008	-0.004	0.011*	-0.001
	(0.010)	(0.010)	(0.007)	(0.007)
White-Asian Gap	0.043	0.043	0.010	0.029**
	(0.032)	(0.027)	(0.015)	(0.012)
Econ. Status Gap	-0.007	-0.002	0.006	-0.004
	(0.006)	(0.006)	(0.007)	(0.007)
Number of Obs.	339,303	372,672	441,540	486,508

Notes: Table A.1 presents two-stage difference-in-differences estimates of the effect of non-chain and in-home chain tutoring center entry on district-grade test scores. The treatment variable is an indicator for whether a school district has received its first non-chain (Columns (1) and (2)) or in-home chain tutoring center (Columns (3) and (4)), respectively. Columns (1) and (3) report impacts on math test scores; Columns (2) and (4) on reading. Each cell is a separate regression. All specifications include district-grade and year fixed effects but no additional controls. The average treatment effect for each regression is calculated using equation (4), omitting period 0. Standard errors are bootstrapped and clustered by school district. Rows report estimates for all students and for subgroups defined by race/ethnicity, economically disadvantaged status, gender, grade level, and within-district test score gaps. Tutoring center locations and chain classifications come from Data Axle. School district boundaries and test score data come from SEDA. The sample contains first-treated and never-treated school districts from 2009–2019 with at least 10 years of observed test score data. For more details on sample construction, see section 5.1. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: Two-Stage DiD with Alternative Weighting and State × Year FEs

	Baseline (1)	Weighted (2)	State x Year FEs (3)	Both (4)
Panel A. Math				
All	0.034***	0.027**	0.037***	0.031***
	(0.010)	(0.011)	(0.008)	(0.010)
White/Asian	0.043***	0.041***	0.044***	0.047***
	(0.009)	(0.010)	(0.009)	(0.009)
Black/Hisp./Nat. Am.	0.021*	0.023	0.036***	0.011
	(0.012)	(0.015)	(0.013)	(0.015)
Non-Econ. Disadvantaged	0.025***	0.030***	0.033***	0.036***
	(0.009)	(0.009)	(0.008)	(0.009)
Econ. Disadvantaged	0.025**	0.013	0.037***	0.018
	(0.010)	(0.013)	(0.009)	(0.014)
Number of Obs.	381,356	381,356	381,356	381,356
Panel B. Reading				
All	0.021***	0.021***	0.017***	0.018**
	(0.006)	(0.008)	(0.006)	(0.008)
White/Asian	0.021***	0.018**	0.018***	0.021***
	(0.007)	(0.008)	(0.007)	(0.007)
Black/Hisp./Nat. Am.	0.022**	0.018	0.020**	0.010
	(0.009)	(0.014)	(0.010)	(0.014)
Non-Econ. Disadvantaged	0.013	0.013	0.015**	0.018**
	(0.008)	(0.009)	(0.007)	(0.007)
Econ. Disadvantaged	0.018**	0.024**	0.017**	0.018*
	(0.008)	(0.012)	(0.007)	(0.010)
Number of Obs.	418,618	418,618	418,618	418,618
Add. Controls	X	X	X	X
Weighted		X		X
State X Year FEs			X	X

Notes: Table A.2 presents the average treatment effect of tutoring chain entry on math and reading test scores under alternative weighting and fixed effects specification. Each cell is a separate regression. All estimates include calendar-year and district-grade fixed effects. Column (1) replicates our baseline result with additional controls from Columns (2) of Tables 3 and 4. Additional controls include grade-level enrollment, grade-level racial shares, district-level unemployment and median income. All district-grades weighted equally, and the specification includes district-grade and year fixed effects. Column (2) weights each district grade by the number of test-takers in the demographic subgroup (e.g. White/Asian test scores outcomes are weighted using the number of White/Asian test-takers). Column(3) includes state by year fixed effects. Column (4) does both. Panel A. presents results for math test scores, panel B. for reading. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3: Alternative Restrictions on Required Years of Observation

Required Years of Obs.	11 years (1)	**10 years** (2)	9 years (3)	7 years (4)	5 years (5)	2 years (6)
All	0.029***	**0.029*** **	0.030***	0.028***	0.028***	0.029***
	(0.011)	**(0.008)**	(0.008)	(0.007)	(0.008)	(0.009)
White/Asian	0.050***	**0.040*** **	0.042***	0.041***	0.042***	0.043***
	(0.011)	**(0.009)**	(0.009)	(0.008)	(0.008)	(0.008)
Black/Hisp./Nat. Am.	0.021	**0.018**	0.022*	0.024**	0.018*	0.018
	(0.015)	**(0.013)**	(0.012)	(0.011)	(0.011)	(0.011)
Non Econ Disadvantaged	0.026**	**0.021** **	0.023**	0.023**	0.024***	0.025***
	(0.011)	**(0.009)**	(0.009)	(0.009)	(0.008)	(0.009)
Econ Disadvantaged	0.025**	**0.018** **	0.016*	0.016	0.014	0.016
	(0.013)	**(0.009)**	(0.009)	(0.010)	(0.011)	(0.010)
Number of Obs.	261,536	**381,356**	417,536	448,108	483,868	499,132

Notes: Table A.3 presents the average treatment effect of tutoring chain entry on math test scores under alternative minimum-observation requirements, replicating the specification of Column (1) of Table 3. Each Column imposes a different threshold for the minimum number of years a district-grade observation must appear in the SEDA test score panel to be included in the sample: 11 years (Column (1)), 10 years (Column (2), bolded, the baseline specification used throughout the paper), 9 years (Column 3), 7 years (Column 4), 5 years (Column 5), and 2 years (Column 6). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4: Results When Defining Chain Exposure with 10-Mile Radius

	2-stage DiD with First Chain		TWFE with All Districts	
	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)
All	0.026**	0.017**	0.003***	0.004***
	(0.011)	(0.008)	(0.001)	(0.001)
White/Asian	0.037***	0.019**	0.005***	0.005***
	(0.010)	(0.008)	(0.001)	(0.001)
Black/Hisp./Nat. Am.	0.012	0.033*	-0.000	0.004***
	(0.019)	(0.017)	(0.001)	(0.001)
Non-Econ. Disadvantaged	0.037***	0.018*	0.004***	0.004***
	(0.013)	(0.009)	(0.001)	(0.001)
Econ. Disadvantaged	-0.002	0.001	-0.000	0.005***
	(0.012)	(0.009)	(0.001)	(0.001)
Number of Obs.	274,377	302,314	461,015	508,907

Notes: Table A.4 presents estimates of the effect of chain tutoring centers on district-grade test scores when defining districts as treated if there is a chain center in a 10-mile radius of the district centroid. Columns (1) and (2) present estimates using our baseline 2-stage difference-in-differences approach, duplicating the specifications from Column (1) of Tables 3 and 4 respectively for math and reading scores. Columns (3) and (4) using a continuous measure of chains within a 10 mile radius per 1,000 children ages 5-17 on math and reading test scores in all school districts, duplicating the specifications from Column (5) of Tables 3 and 4. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.