NBER WORKING PAPER SERIES

PATTERNS, DETERMINANTS, AND CONSEQUENCES OF ABILITY TRACKING: EVIDENCE FROM TEXAS PUBLIC SCHOOLS

Kate Antonovics
Sandra E. Black
Julie Berry Cullen
Akiva Yonah Meiselman

Working Paper 30370
http://www.nber.org/papers/w30370

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2022 Revised August 2023

The authors are grateful to Mindie Hsu, Eli Mogel, Anjali Priya, Kelly Wang, and Sonia Yan for excellent research assistance. We thank Sarah Cohodes, Hilary Hoynes, and Sean Reardon for helpful suggestions. The conclusions of this research do not necessarily reflect the opinion or official position of the Texas Education Research Center (ERC), the Texas Education Agency, or the State of Texas. All output created using data from the Texas ERC, including tables and graphs, have been reviewed by Texas ERC staff to ensure FERPA compliance. This work was partially supported by the Spencer Foundation Small Grant Program and the Research Council of Norway through its Centres of Excellence Scheme, FAIR project No. 262675. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Patterns, Determinants, and Consequences of Ability Tracking: Evidence from Texas Public Schools

Kate Antonovics, Sandra E. Black, Julie Berry Cullen, and Akiva Yonah Meiselman

NBER Working Paper No. 30370
August 2022 Revised August 2023
JEL No. H75,I21,I24,I28

ABSTRACT

Little is known about the pervasiveness or determinants of within-school ability tracking in the US. To fill this gap, we use detailed administrative data to estimate the extent of tracking in Texas public schools for grades 4 through 8 over the years 2011-2019. Strikingly, we find that ability tracking across classes within schools overwhelms sorting by ability across districts and schools, as well as sorting by race/ethnicity or economic disadvantage. We also examine how schools operationalize tracking as well as the local characteristics that predict tracking. Finally, we explore how exposure to tracking (and the bundle of associated practices) relates to achievement gains, finding that, on average, tracking increases inequality by slightly improving test scores of higher-achieving students without harming those of lower-achieving students.

Kate Antonovics
Department of Economics - 0508
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
kantonov@ucsd.edu

Julie Berry Cullen
Department of Economics - 0508
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
and NBER
jbcullen@ucsd.edu

Sandra E. Black
Department of Economics and
School of International and Public Affairs
Columbia University
1022 International Affairs Building
420 West 118th Street
New York, NY 10027
and IZA
and also NBER
sblack@columbia.edu

Akiva Yonah Meiselman
Research Improving People's Lives
Providence, RI 02903
yonah.meiselman@gmail.com

A data appendix is available at http://www.nber.org/data-appendix/w30370

1. Introduction

A major goal of public education is to provide students with opportunities for economic and social mobility. At the same time, schools often assign students to classrooms based on academic ability, effectively mimicking the very stratification that public education is intended to combat. Proponents of ability tracking—the sorting of students across classes within school based on ability—argue that it is a low-cost tool to improve learning since instruction is more effective when students are segregated by ability, while opponents argue that tracking exacerbates initial differences in opportunities without strong evidence of efficacy.[2]

In fact, existing research has not come to a consensus on the efficacy of tracking across classes in elementary and secondary schools. Early research from economists and sociologists suggested that tracking benefitted high-ability students at a cost to low-ability students, leading to a pushback against tracking in the US.[3] More recent research has questioned the validity of the early studies and employed alternative identification strategies. Yet these newer studies have yielded mixed evidence, with some uncovering evidence of negative effects of tracking on low-ability students (e.g., Bacher-Hicks and Avery 2018; Fu and Mehta 2018) and others finding the opposite (e.g., Collins and Gan 2013).[4]

Even more basic, relatively little is known about the scope and nature of tracking in the US. This is in large part because the ways by which students are assigned to classrooms according to ability are often informal, in contrast to systems common in other countries that stream students to different schools or programs of study. National surveys of school principals suggest that tracking by ability across classes is prevalent in the US. These reveal that on the order of one-quarter of 4th graders and three-quarters of 8th graders are served in schools that track, and that the US is an outlier—along with the Ireland and the UK—in its reliance on this form of student sorting.[5]

In this paper, we take advantage of detailed administrative data from Texas—a state with 10% of the school-aged population in the US, covering more than 1,200 districts and 8,800

[2] There are numerous ways in which students are grouped by ability over the course of their schooling. Following Loveless (2013), we use the term "tracking" to refer to the sorting of students across classes within the same school.
[3] See Betts (2011) for a comprehensive review.
[4] Some of the most compelling research has been done in developing country contexts, where students are randomly assigned to tracked or untracked regimes. In this case, evidence suggests that student performance increases for all students under the tracking regime (e.g., Duflo et al. 2011).
[5] The sources for these statistics are the 2015 National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMMS). For more details, see Appendix A.

schools—to quantify the degree to which students are grouped by ability across classes in public schools.[6] Using data from 2011 to 2019, we calculate two data-driven measures of tracking for grades 4 to 8 across math classes according to prior math scores.[7] The first is an R-squared statistic capturing how much of the variation in prior math test scores can be explained by current math class assignments (Lefgren 2004), and the second uses simulations to estimate how sorted students are relative to the maximum possible given the class size and student achievement distributions (Hellerstein et al. 2011). The first "absolute" tracking measure embeds the role of class size choices, while the second "relative" measure controls for these. Relative to survey-based measures, our measures have the advantages of being comparable across schools and reflecting not only the incidence but also the intensity of tracking. Importantly, our measures also capture all means by which students are sorted across classes, ranging from purposeful assignment for curricular or instructional differentiation to the unintended byproduct of other factors affecting class assignments, such as parental preferences for certain teachers.[8]

We use our data-driven measures to provide new insights into the nature and determinants of tracking in Texas. We answer questions such as, how important is within-school tracking in the grand scheme of student sorting? What are the explicit and implicit mechanisms by which students are tracked? Which schools and districts track students to a greater degree? Finally, we consider the impact of exposure to more tracked regimes on future achievement at different parts of the initial statewide achievement distribution.

Our first striking finding is that tracking by ability within schools overwhelms any sorting by ability that takes place across schools. A popular perception in the US is that much of the sorting takes place across districts and schools, since school assignment is based primarily on residential location. In fact, only 9% and 17% of the variation in prior scores (within grade-years) is explained by districts and schools, respectively, while 44% is explained by classes. Our results also suggest that within-school sorting based on prior test scores is far greater than within-school sorting based on race/ethnicity and SES. In addition, we find substantial variation in tracking across grades and schools. Consistent with national survey data, we find that middle

[6] The sources for the population and school statistics are De Brey et al. (2021) and Texas Education Agency (2020).
[7] We choose to focus on math given the evidence on high returns to math achievement and coursework (e.g., Goodman 2019).
[8] Like other across-class tracking measures, our measures miss the extent to which students are sorted into different ability groups within the same classroom.

school grades track more than elementary schools. And, while the average elementary (middle) school student in our sample is in a school that realizes about 12% (42%) of its potential to track students across classes, this ranges from no tracking at the 5th percentile (both elementary and middle schools) to 47% (66%) at the 95th percentile.

Schools can facilitate tracking in a variety of ways. When we examine the decisions that are most predictive of tracking, we find that schools appear to operationalize tracking through more aggressive classification of students into categories such as gifted and disabled, as well as through differentiating the math curriculum to offer more remedial and advanced options. This is true even after controlling extensively for the distribution of student ability.

In terms of which schools track, we find that the most important determinant is heterogeneity in student ability. In school-grade cohorts with more heterogeneity, as measured by the standard deviation of prior test scores, we see substantially more tracking. Interestingly, the racial composition of the school is unrelated to the level of tracking once we control for the distribution of student ability. Other findings are that tracking is less prevalent in charter schools and in districts with larger private school enrollment shares, and uncorrelated with how Democratic the county's residents vote in presidential elections.

To explore the implications of tracking, we consider how exposure to tracking across cohorts within districts relates to student test score growth across the distribution of initial achievement. To do so, we map students' positions in the statewide test score distribution in third grade to their positions in the test score distribution five years later. We find that for students at the bottom of the test score distribution, exposure to tracking (and the associated bundle of practices) is not generally related to future test score growth. For those initially at the top, however, exposure to more tracking is on average beneficial. For example, our results suggest that a one standard deviation increase in exposure to middle-school tracking would lead to a 0.6 percentile increase in predicted test scores 5 years after 3rd grade for students initially at the 75th percentile. These findings are consistent with tracking slightly aggravating inequities in educational outcomes, but primarily by benefitting those already at the top.

To examine possible mechanisms for the effects on achievement growth, we use a similar empirical strategy to examine how tracking relates to the level of the math curriculum, average class size, and peer quality experienced by students at different points in the initial test score distribution. Tracking in elementary and middle school grades reduces the likelihood that

initially low-achieving students are in math courses beyond grade 8 five years after grade 3, while tracking in middle school grades increases this likelihood for high-achieving students. Not surprisingly, students who are exposed to more tracking face more inequality in peer achievement across lower- and higher-achieving classes, and less variability within classes. In addition, we find that class sizes are on average smaller for students exposed to more tracking, especially for students at the bottom of the initial test score distribution.[9]

Our study contributes to several literatures. The first is related to the measurement of tracking. Most prior studies that have used similar data-driven approaches have focused on a single school district (e.g., Collins and Gan, 2013, using data from Dallas, and Lefgren, 2004, using data from Chicago) or a limited number of school districts (Kalogrides and Loeb, 2013, using three large urban school districts). Our paper builds upon this work by measuring tracking for a larger and more diverse population. Dalane and Marcotte (2020) and Clotfelter et al. (2021) also use student-level administrative data to examine tracking for a large, diverse population (in North Carolina), but their work focuses on sorting across classrooms by socioeconomic status.[10]

The second is the literature studying the determinants of tracking. Epple, Newlon, and Romano (2002) develop a theoretical model of education markets where public schools track to retain higher-income, higher-ability students. In support of this prediction, Figlio and Page (2002) find that when a school introduces tracking, the share of students at the school that is eligible for free lunch falls. Our finding that more tracking is correlated with lower private school shares might thus be expected in equilibrium.

Finally, we contribute to the literature studying how tracking affects educational opportunity across the ability distribution. In addition to the work mentioned at the outset on generic ability tracking, there are several recent studies that exploit rules or policy changes that determine placement in specialized high- or low-achieving classes for identification.[11] For example, Card and Giuliano (2016) and Cohodes (2020) use regression discontinuity designs and

[9] Under tracking, others have found evidence of adjustments to class sizes and teacher quality that in some cases reinforce and in others compensate for differences in peer quality (e.g., Bacher-Hicks and Avery 2018; Betts and Shkolnik 2000; Rees, Brewer, and Argys 2000).

[10] There are also several studies that quantify the degree to which ability sorting across classes introduces bias in estimates of teacher value added. This includes work by Aaronson et al. (2007), Alzen and Domingue (2013), Clotfelter et al. (2006), Dieterle et al. (2014), and Horvath (2015).

[11] Another strand of empirical literature has exploited policy variation in streaming across schools or programs that is more common in European countries (e.g., Bauer and Riphahn 2006, Clark and Del Bono 2016, Dustmann et al. 2017, and Hanushek and Woessmann 2006).

find that students granted access to high-achiever classes benefit, with no evidence of negative effects on other students. Ballis and Heath (2021) and Cortes and Goodman (2014) find low-achieving students benefit from placement in special education and remedial classes, respectively, despite exposure to lower-ability peers. In our case, we examine the average relationship between exposure to more- and less-tracked regimes and test score mobility across Texas, relying on across-cohort variation for identification. Our approach is most similar to recent work by Reardon (2019), which uses administrative test score data to document patterns of achievement gains across grades for US school districts.

The rest of the paper unfolds as follows. In the next section, we discuss the data and methods we use to quantify tracking. Section 3 then examines the incidence of tracking in Texas and the programmatic choices that underlie the observed sorting. In sections 4 and 5, we move on to explore the determinants and consequences of tracking for different types of students. Section 6 offers a brief concluding discussion.

2. Data and Methodology

2.1 Administrative Data and Sample

We rely on administrative data from the Texas Education Agency (TEA) available through the Texas Education Research Center. These data cover the universe of public elementary and secondary school students in Texas and enable us to follow students over time. While earlier data are available, we only observe classroom assignments beginning with 2011 (i.e., the 2010-11 school year). For students, we have a limited number of demographic characteristics, along with enrollment and coursework by school and term, and achievement as measured by standardized test scores. To supplement these restricted-use data, we merge information on school and district characteristics from publicly available annual reports from the TEA.

As a proxy for student ability, we use test scores from standardized mathematics tests taken in the prior year. Between 2003 and 2011, the Texas Assessment of Knowledge and Skills (TAKS) was the primary statewide assessment program. TAKS was designed to measure performance on the state-mandated curriculum and involved the administration of standardized tests in grades 3 through 11. From 2012 on, the state switched to the State of Texas Assessments of Academic Readiness (STAAR) program, adjusting standards and replacing grade-specific

assessments with course-specific end-of-course exams for high school students and middle school students taking high school courses. This switch acknowledges that curriculum differentiation in higher grades goes beyond teaching the same material at different levels.

Key barriers to measuring tracking past grade 8 are that the end-of-course scores are not comparable across courses and that high school students often take no math course at all in a given term. We are also unable to consider grades before grade 4, since prior-year test scores are not available. Thus, we analyze tracking in grades 4 through 8. For students in these grades, we start with their prior-year math scale scores from the grade-specific assessments. These scores are almost always available for continuing students, and the vertical scales are meant to be comparable across grades and years within the two testing regimes.[12] We convert students' prior-year math scale scores to z-scores by subtracting the statewide mean and dividing by the statewide standard deviation for the relevant grade and year.[13]

With prior math achievement in hand, the next step is to identify students' math classes. We start with students enrolled at a given school at the start of the fall term. In most cases, it is straightforward from the transcript record to identify their math classes. In some cases, schools use generic course titles for all subjects (such as "Grade 4"), or students take multiple math courses in a single term.[14] In the former case, the same students are typically grouped together for all subjects, and we select one representative class for them. In the latter, we choose the math course that enrolls the largest number of same-grade peers. Enrolled students who have neither math nor generic course transcript records are not allocated to a class.

Thus, the sample of students we use to estimate tracking is the set of enrolled students with non-missing prior scores for whom we can identify a focal math class.[15] We include all

[12] Importantly, there is very little bunching at the top of the distribution, with less than 1% of students achieving the maximum possible score for their test.

[13] Prior-year scores are normalized by the statewide distribution for the prior grade even for students who are retained or otherwise off track. For example, students retained in grade 4 have their prior-year grade 4 scale scores normalized using the prior-year grade 3 distribution, matching the normalization used for their on-track peers with prior-year grade 3 scale scores.

[14] These cases are rare: 1.0% of students are in generic courses (mostly in grades 4-5), and 0.3% are taking multiple math courses (mostly in grades 7-8).

[15] Table 1 shows that the shares of students without a focal math course and missing prior test scores range from 4-10% and 6-7% across grades, respectively. Test scores may be missing due to student absence or migration (across states or sectors), or due to test-taking exemptions. Exemptions for students receiving special education services were more lenient up through the 2013-14 school year, after which the US Department of Education decided that assessments based on modified standards would no longer count toward accountability. And, under the STAAR regime, students take an end-of-course assessment rather than the grade-level math assessment if they are receiving instruction in a high school level course (e.g., algebra or geometry).

school-grade-years from 2011 to 2019 with at least two classes with two or more students. Table 1 presents (student-weighted) summary statistics by grade for this sample, which represents over 4,000 elementary and 2,000 middle schools across 1,000 districts. These schools serve 96% of students enrolled in grades 4-8 in Texas public schools over our study period.

2.2 Measurement of Tracking

We build on data-driven measures developed in prior studies to quantify tracking by ability across classes. Our first measure is an "absolute" measure that embeds any role of the class size distribution, and the second is a "relative" measure that captures how sorted students are conditional on that distribution. Both measures are defined at the school-grade-year level.

As our "absolute" measure of tracking (ρ), we borrow the measure used by Lefgren (2004) as part of an instrumental variables strategy to estimate peer effects in the Chicago Public Schools. Lefgren estimates the relationship between students' prior-year test scores and indicators for the specific classes in which they are enrolled in the current year. His proxy for the degree of tracking is the R-squared from this regression, which reflects how much a student's own achievement can be predicted by the achievement of the student's classmates. If students are randomly assigned to classes within a given school and grade, average ability will not vary by class and the class indicators will have little explanatory power for prior test scores; the measure will then be close to zero. Alternatively, if students are grouped strictly by ability, the class indicators will strongly predict prior test scores and the R-squared will be high. Importantly, although it is sensitive to the number of classes students are spread across, this R-squared measure is mechanically invariant to changes in the variance of student achievement. Another nice feature of this measure is that we are able to test whether it is statistically different from zero—that is, whether we can reject the null hypothesis of no tracking—using the F-statistic.[16]

Since class sizes may be determined by resource levels or policies unrelated to tracking, our second "relative" measure attempts to isolate tracking independent of the class size distribution. To do this, we take the class size and student ability distributions at the school-grade-year level as given and calculate the fraction of potential sorting that is realized. These adjustments could matter if class size constraints and higher-order aspects of the ability distribution limit the ability of schools to sort students, even when they may want to. For

[16] See Appendix B for more details on both of our measures and their properties.

example, compared to an otherwise identical cohort, one that is spread across two classes rather than three has less scope for sorting. And, compared to a cohort with the same variance in prior achievement, one that is characterized by three ability types cannot be sorted as strongly across two classes as one characterized by two ability types. We use simulations to account for these factors in a nonparametric way.

Our relative tracking measure is equal to the ratio of the observed deviation of the R-squared from what would be expected under random assignment ($\rho^{ra,\mu}$) to the expected deviation under strict tracking:[17]

$$\rho^{rel} = \frac{\rho - \rho^{ra,\mu}}{\rho^{strict,\mu} - \rho^{ra,\mu}}$$

This can loosely be interpreted as the share of potential tracking that is realized.[18] The expected R-squared from regressing prior scores on class indicators under random assignment across permutations, $\rho^{ra,\mu}$, is readily calculated as a simple function of the numbers of students and classes.[19] To simulate the expected R-squared under strict tracking, $\rho^{strict,\mu}$, we rank students based on prior-year test scores and then, taking the number and sizes of classes as given, repeatedly (i.e., 1,000 times) randomly order the classes and assign students to classes with the top-scoring students assigned first. We then calculate the mean of the estimated R-squared from regressing prior scores on class indicators across permutations.

Which of the two tracking measures is of greater interest depends on the question. The absolute measure is most informative about the overall degree to which students are sorted. The relative measure is useful when trying to parse out tracking that is independent of class sizes, which may be driven by other considerations that have their own impacts on outcomes.

3. Scope and Nature of Tracking

In this section, we first present our findings on the how the degree of sorting by ability across classes within a school compares to sorting at other levels, such as across districts and

[17] This measure is similar in spirit to the measure of "effective network isolation" used by Hellerstein et al. (2011).
[18] The interpretation is loose since the ratio can be less than zero when the actual measure is below the expected value under random assignment, and greater than one when the actual measure is above the expected value under strict tracking.
[19] We use the mean and standard deviation of the distribution of the R-squared under random assignment to construct an alternative finite sample test for whether or not the observed degree of tracking is statistically significant. We show in Appendix B that inference from this alternative strategy corresponds closely to the more traditional F-test.

across schools within districts, and to sorting on other dimensions. We move on to discuss the magnitudes and variation in ability tracking overall and by grade level. We then explore the ways that schools operationalize tracking, such as through curricular differentiation and the classification of students in categories that have special needs.

3.1 Comparative Scope of Ability Tracking

Because school assignment in the US is based primarily on residential location, it is possible that substantial sorting by ability has already taken place across districts and schools, limiting the capacity for tracking across classrooms within schools. As initial evidence against this claim, columns 1-3 of Table 2 document the share of the variation in student prior-year math test scores that can be explained by different levels of fixed effects. The R-squared statistics from sequentially regressing individual z-scores on district, school, and class indicators shown in the top row reveal that only 9% and 17% of the variation in prior scores (within grade-years) is explained by districts and schools, respectively, while 44% is explained by classes.

Figure 1 provides another perspective on the importance of within-school tracking relative to sorting across schools within districts. It shows the distributions of the absolute and relative tracking measures that capture sorting by prior achievement across classes within school-grade-years (grey bars), alongside the distributions of these same measures when calculated based on the sorting of students across schools within district-grade-years (black bars). As is clear from the limited overlap in the distributions far above the no-tracking benchmarks of zero, across-school sorting by ability – after residential and school choices are made and before students arrive in the classroom – is much lower than sorting across classes within schools.

While sorting within schools by ability overwhelms any sorting across schools or districts, this is not true for sorting by race/ethnicity or by economic disadvantage.[20] The remaining columns in Table 2 calculate the series of R-squared statistics using an indicator for Black or Hispanic (columns 4-6) and an indicator for low income (columns 7-9) as the dependent variables instead of test scores.[21] The statistics in the top row reveal that students are more sorted

[20] This is consistent with work by Kalogrides and Loeb (2013), which uses data from 3 large urban school districts and shows that sorting by race and poverty status within schools is less than sorting across schools.
[21] Though these outcomes are binary, the R-squared is meaningful since the regression model is fully saturated. When the model includes only mutually exclusive indicators, the R-squared measures the share of the variation in the outcome explained by differences in means across the groups (i.e., districts, schools, or classes).

on these dimensions across districts and schools than they are by prior achievement, and sorting across classes within schools plays a smaller role. District, school, and class fixed effects account for 26%, 34%, and 40% of the variation in Black or Hispanic status, and 19%, 30%, and 37% of the variation in low-income status.

Analogous to Figure 1, Figure 2 shows the distributions of the absolute and relative tracking measures when these are calculated to capture sorting by race/ethnicity or economic disadvantage across classes within school-grade-years (grey bars) and across schools within district-grade-years (black bars). It is apparent that, across classes within schools, there is less sorting by race/ethnicity and low-income status than by prior test scores. The distributions of our within-school absolute and relative tracking measures when indicators for Black or Hispanic or low income are used in place of prior achievement are much more tightly clustered around the no-tracking benchmarks.

These findings may be surprising, given that some sorting on these demographic dimensions would be expected if schools are tracking by ability due to their correlations with test scores. Moving from the 25th to the 75th percentile in the statewide test score distribution raises scores by one standard deviation, while the shares of students classified as Black and Hispanic or low-income decrease by less than 20 percentage points.[22] Thus, sorting by test scores only weakly induces sorting by these demographics.

3.2 Scope of Ability Tracking Overall and by Grade Level

Returning to our central measures of within-school tracking by ability across classes (shown in Figure 1, grey bars), the mean level of absolute tracking across grades 4-8 is 0.32, and the standard deviation is 0.21. This implies that the average student is in a cohort where class assignments explain 32% of the variation in prior scores. Viewing these as continuous measures of the degree of tracking, values above 0.15 are almost always statistically significantly different from zero (See Appendix B). The mean level of relative tracking is 0.30, and the standard deviation is 0.24. Using our loose interpretation of relative tracking, this suggests that on average

[22] Appendix Figure C1 shows race/ethnicity and low-income status breakdowns across the statewide achievement distribution in grade 3. The shares Black and Hispanic and low-income fall steadily moving from lower to higher percentiles. Within school-grade-year cohorts, Black and Hispanic students' and low-income students' prior-year math scores are about 0.4 standard deviations below their peers on average. This is the same test score gap as moving from the 50th to the 65th percentile of the statewide test score distribution.

30% of potential sorting by prior achievement across classes is realized by actual class assignments. The correlation between our two tracking measures is very high at 0.99.

To get a sense of what these magnitudes mean for the classroom, Figure 3 relates our tracking measures to the standard deviation of prior scores within classes, which is 0.74 on average. In the relevant ranges (where most of the densities are), the relationships between both tracking measures and the standard deviation are approximately linear with slopes of about -0.5. That is, an increase of 0.10 in either tracking measure is associated with a decline of -0.05 standard deviations in the dispersion of classroom peer achievement. Considering the distribution of absolute tracking, the average standard deviation of classroom peer achievement is 0.88 for the school-grade-year cohort at the 5th percentile of tracking and 0.59 for the cohort at the 95th percentile. The numbers are the same when we use the distribution of relative tracking instead.

Figure 4 shows the distribution of the tracking measures by grade, revealing that the extent of tracking across math classes increases markedly as students move from the elementary to the middle school grades.[23] This pattern, also observed in column 3 of Table 2 moving across panels, helps to explain the bimodality observed for within-school tracking in Figure 1. It is also expected, since sorting by ability will rise as students begin to take courses that are differentiated not only by level of difficulty and pace but also by subject content. Notably, the patterns in Table 2 reveal that sorting across classes by race/ethnicity and low-income status do not increase in the same way in the middle school grades. The share of the variation in those characteristics that is explained by class indicators actually declines somewhat (columns 6 and 9 across panels), in part driven by reduced across-school sorting (columns 5 and 8 across panels) as elementary schools feed into a smaller number of middle schools.

In addition to the differences across grade levels, Figure 4 reveals substantial variation in tracking within grade levels. The fraction of potential tracking realized ranges from none at the 5th percentile to 47% at the 95th percentile for elementary school students, and from none at the 5th percentile to 66% at the 95th percentile for middle school students. Moving from the 5th to 95th percentiles of relative tracking, the average standard deviation of prior test scores within classes falls from 0.89 to 0.69 for elementary school students and from 0.83 to 0.56 for middle school students.

[23] Appendix Figure C2 shows that grade configuration also matters, in that middle school cohorts served in schools that also have elementary grades are less tracked. Figure C3 shows that tracking increases slightly across years.

While the extent of math tracking increases with grade level, it may be more likely to spill over to tracking in other subjects in earlier grades, since elementary school students are more likely to be grouped together with the same teacher for the entire day. To examine this, we continue to use prior math scores as the proxy for student ability but recalculate our tracking measures for other subjects: English language arts/reading, science, and social studies classes. These measures then capture how sorted students are according to math ability in non-math classes and are readily comparable to our baseline measures. Table 3 shows that the correlations in tracking between math and other core subjects range from 0.85 to 0.90 in the elementary grades and from 0.52 to 0.68 in the middle school grades.[24] These results suggest that students more tracked in math classes are also more tracked in other classes. Further, any given degree of sorting across math classes translates to a greater degree of sorting throughout the school day in the elementary grades.

3.3 Nature of Tracking

To provide a sense of how coordinated and purposeful tracking policy is, we next examine how harmonized tracking is across schools within a district. Specifically, we regress our school-grade-year tracking measures successively on district, district-grade, and district-grade-year fixed effects. Across all school-grade-year cells, the top panel in Table 4 reveals that 83% of the variation in our absolute tracking measure is explained by district-grade-year fixed effects. When we focus exclusively on larger districts (with at least 6 schools for every grade across all years), the fraction of the variation accounted for by district-grade-year fixed effects falls to 73%, which is still substantial. When we report results separately by grade level in the bottom panels of Table 4, we find the district plays a more important role in middle school, where district-grade-year fixed effects account for 74% (vs. 59%) of the variation in tracking.[25] Taken together, these findings suggest that the district plays a substantial role in setting policies and practices that affect tracking.

We then examine the different ways schools operationalize the sorting of students across math classes within a school. The assignment of students to classrooms based on ability could

[24] These correlations are likely understated due to measurement error. Appendix Figures C4 and C5 show the distributions of the absolute and relative tracking measures for all four core subjects for visual comparisons.
[25] Further evidence suggesting that tracking practices are intentional is persistence across time. Almost 80% of the variation in the tracking measures at the school-grade-year level is explained by school-grade fixed effects.

arise from numerous behaviors and policies by parents and administrators. It might be inadvertent on the part of the school, such as if high-SES parents successfully push for specific teacher assignments, or purposeful, such as if administrators use achievement as a factor in class and course assignments. To facilitate tracking, schools or districts could adjust class sizes or offer more advanced or remedial course offerings. There are also relevant state policies regarding special student populations, such as gifted students, students with disabilities, and English learners, and the classification of students into these categories could facilitate tracking. An advantage of our tracking measures is that they embed all these factors, while a disadvantage is that it is challenging to decompose them.

As a step toward identifying the factors that give rise to tracking, we regress our school-grade-year absolute tracking measure on school-grade-year characteristics that are intended to capture programming choices that could lead to the segregation of low- and high-achieving students. Table 5 presents the results.[26] Across all specifications, we include controls for the mean and standard deviation of the cohort's prior-year math scores, as an effort to hold the distribution of ability constant. To compare cohorts that are similar on other key dimensions including fiscal capacity, we also include controls for grade level, grades served at the school, cohort and district size, district property wealth, type of locale, and year. Observations are weighted by cohort enrollment, and standard errors are clustered at the district level.

In the first column of Table 5, we include variables related to special needs populations and instructional settings. For students with limited English proficiency (LEP), we include the overall share as well as the share receiving instruction in other core subjects in segregated settings (i.e., English as a second language (ESL) content-based and bilingual non-two-way instructional models), as opposed to settings where they are integrated with other students (i.e., ESL pull-out and bilingual two-way models). We also include the overall shares of students classified with physical and non-physical disabilities, as well as the share served in restricted settings where less than half of the day is spent in general education classrooms. Finally, we control for the fraction classified as gifted. Importantly, other than for the classification of students as LEP and as having physical disabilities, which are mostly formulaic and objective, schools have substantial discretion in classifying students. For classified students, schools have discretion in determining the services they receive and the learning environments in which those

[26] In results not shown, we repeat this analysis for our relative tracking measure, yielding similar results.

are provided.[27] Rather than being tags for specific programs, as might be the case in other countries, these classifications follow students across subjects and often across grades.

The results in column 1 of Table 5 reveal that similarly able cohorts with more students classified in special needs categories and segregated for instruction are also more tracked by ability across classes. Focusing first on the shares classified in different categories, the positive coefficients on the shares LEP and with physical disabilities most likely reflect the impact of student case-mix. However, the positive links between tracking and the share classified with non-physical disabilities, which are dominated by emotional and learning disabilities, and the share identified as gifted, likely reflect differences in practices as well, since classification is much more subjective. Though more aggressively classifying students as disabled and gifted is associated with more tracking, the implied magnitudes of the point estimates are small. For example, the point estimate of 0.314 in the fourth row implies that a 1 standard deviation (0.033) greater share with non-physical disabilities is associated with an absolute tracking measure that is 0.01 greater, which is 0.05 standard deviations.

We turn next to the associated instructional settings. Since the potential pros and cons related to serving students with special needs in self-contained vs. general education classes are very much the same as those surrounding segregating students by ability, we would expect the choices to be related. There is also a potential direct tie if students grouped together in self-contained classes are more homogenous in terms of ability. While the results in column 1 of Table 5 reveal no significant relationship between tracking and serving LEP students in more isolated settings, serving students with disabilities in more restrictive settings is associated with cohorts being more tracked.

In column 2, we add controls for curricular differentiation and instructional resources. Our measure of curricular differentiation captures the dispersion of students across different math courses and is equal to one minus the Herfindahl index calculated based on students' course titles. School-grade-years with only one course title (e.g., "grade 4 math") have a value of 0, while those with several math course titles have higher values. Not surprisingly, math curricular differentiation is associated with greater tracking, with a 1 standard deviation increase mapping to tracking that is higher by 0.16 standard deviations. We also find that cohorts that are more tracked have access to greater instructional resources, including more experienced teachers

[27] Summary statistics for these variables for all grades combined and separately by grade are provided in Table 1.

and smaller classes. As might be expected, adding these additional programmatic controls moderates some of the relationships observed in column 1. Most notably, there is no longer a statistically significant relationship between physical disability rates and tracking, and a higher share of LEP students in segregated settings predicts the cohort overall is less tracked than it otherwise would be.

The remaining columns in Table 5 add successive controls to test sensitivity of the descriptive relationships. These include controls for the tails of the student prior achievement distribution (column 3), district-by-grade fixed effects (column 4), and school-by-grade fixed effects (column 5). Focusing on the results in column 5, which isolate within school-grade variation and control flexibly for cohort ability, we find that, though the other point estimates shrink and are no longer significant, curricular differentiation and resources continue to play the same roles, and more-tracked cohorts are still more often classified in the most subjective special needs categories. Overall, the results in this section highlight that our measures of tracking reflect bundles of instructional policies and practices.

4. Determinants of Tracking

Different schools and districts are likely to perceive the possible equity-efficiency tradeoffs differently, depending on their constituencies. For example, schools serving students with wide disparities in ability might see more instructional benefits to tracking. Research also suggests that parents of high-achieving children (who also tend to be high SES) disproportionately favor tracking (e.g., Figlio and Page 2002). And, on the ideological spectrum, liberals may be less likely than conservatives to support tracking if disadvantaged students do not benefit and achievement gaps increase. In this section, we examine who tracks.

We begin by examining geographic patterns in tracking across Texas. The maps in Figure 5 show district-level variation in (absolute) tracking for the elementary grades in the top panel and middle school grades in the bottom panel. The systematic increase in the level of tracking in the later grades is immediately apparent from comparing the two panels. Otherwise, within each panel, it is striking that there are no noticeable patterns in the level of tracking across more and less densely populated areas, or more and less advantaged areas.[28]

[28] For reference, Appendix Figure C6 shows district-level variation in population density and initial math achievement levels in grade 3, which serves as a proxy for socioeconomic advantage.

We next assess the role of local characteristics on tracking decisions using a regression framework. Table 6 presents the results from regressing our school-grade-year absolute tracking measure on various school, district, and county characteristics.[29] All specifications condition on cohort grade and size, school grade composition, district size and property wealth, type of locale, and year. Observations are weighted by school-grade-year enrollment, and standard errors are clustered at the district level. Columns 1 to 3 show results as covariates are added to capture the student ability distribution, schooling landscape, and local ideology, while the remaining columns show sensitivity to controlling more flexibly for the distribution of cohort prior achievement (added in column 4), district-by-grade fixed effects (added in column 5), and school-by-grade fixed effects (added in column 6).

Consistent with previous studies, column 1 shows that tracking is positively and statistically significantly associated with mean prior test scores, implying that schools that serve high-achieving students tend to track more. However, once we control for the variability of test scores as measured by the standard deviation of prior test scores within a school-grade-year, the coefficient on the mean flips sign (column 2). The standard deviation of prior test scores is itself a positive and statistically significant predictor of tracking.[30] The relationship between tracking and the mean and standard deviation of a cohort's lagged scores becomes less pronounced in columns 4 to 6, as these columns also include controls for lagged math test score percentiles to control more flexibly for student ability. Interestingly, whether we condition on these more flexible controls or not, we find little relationship between student demographics—as proxied by the race/ethnicity composition and shares of students who are low income and limited English proficient—and the degree of tracking.[31]

Focusing on variables related to school type, the results in Table 6 suggest tracking is higher at magnet schools (by .01-.02) and lower at charter schools (by .14-.15). Both types of schools are open to students across school attendance boundaries. Magnet schools focus on specific themes, such as technology or performing arts, and integrate those themes into the core coursework. Though magnets are often designed with the goal of integrating students who may

[29] In results not shown, we find qualitatively similar results for the relative measure.
[30] Recall that the standard deviation is not mechanically related to our measure of tracking, suggesting that the perceived net benefits of tracking increase with the heterogeneity of student ability or that tracking attracts more heterogeneous students.
[31] Though limited English proficiency is positively predictive in Table 5, that is true only conditional on the share that is LEP and served in self-contained settings, which carries the opposite sign with a similar magnitude.

be segregated residentially, we find magnets do more within school sorting across classes than traditional public schools. The opposite finding for charter schools is consistent with evidence from other states that students attending these schools are more evenly distributed across classes compared to traditional public schools (Berends and Donaldson, 2016). Though theoretically tracking might attract or repel students and respond to competitive pressure, we find that higher tracking is associated with a lower district private school share.

With respect to ideology, we find no evidence that an area's political views, as proxied by the county's average Democratic vote share across the 2000-2016 presidential elections, predicts tracking. The negative sign of the point estimate, however, is consistent with the expectation that liberal areas might be less supportive of tracking.[32]

5. Implications of Tracking

Given the prevalence of tracking, a fundamental question is how it affects student academic progress, and how this varies across the achievement distribution. We also want to understand how tracking impacts the educational environment for students of differing ability levels. For these questions, we take a longitudinal perspective and follow cohorts over time.

We limit our longitudinal sample to students in our tracking sample in grade 4 (the first grade for which we have tracking measures) between 2011 and 2015, with non-missing grade 3 math scores from the prior year. For distributional analyses, we characterize these students by their percentiles in the year-specific statewide grade 3 math test score distribution. We then follow them for up to 5 years after grade 3 (which, for most students, is grade 8). A limitation is that we can only observe students as long as they remain enrolled in the Texas Public Schools. By 5 years out, 11% of the overall sample has left the Texas Public School system, and leave rates fall with initial achievement, declining from 12% for students starting at the 25th percentile to 9% for students at the 75th percentile.[33] These leave rates are not differential by exposure to tracking, which should allay concerns about possible attrition biases.

In each subsequent year that students are enrolled, we observe their campus and grade, and thus the level of tracking they experience. For their math classes, we observe the grade level

[32] For the related question of the allocation of students by race across schools, more Democratic school boards are found to adjust school catchment areas to reduce segregation (Macartney and Singleton 2018).
[33] Appendix Figure C7 shows the share enrolled 4 and 5 years out across the distribution of grade 3 test scores.

of the subject, as well as class size and peer quality, where peer quality is proxied by the average initial math test z-scores of classmates based on grade 3 for peers who are part of the longitudinal sample, and the earliest grade available for those who are not. We also observe students' current math test scores, which we convert to statewide percentiles by grade 3 cohort and year. Scores are rarely missing for enrolled students 4 years out, when most are in grade 7, but are frequently missing at the top of the distribution 5 years out, when most are in grade 8.[34] The reason is that these students are taking high-school level courses that have course-specific exams in lieu of grade-level exams. When current scores are missing for enrolled students, we fill in using their most recent available percentile score, which is usually from the prior year. Thus, 5-year-out positions at the top are often in fact 4-year-out positions.

5.1 Achievement Mobility

To examine achievement mobility over time, we follow Reardon (2019) and relate a child's initial position in the test score distribution (in this case, grade 3) to their own position in the test score distribution several years later.[35] Figure 6 shows the relationship between students' initial positions and their percentile ranks in the test score distribution 4 and 5 years later. The relationship is shown separately for students in school-cohorts with above- and below-median exposure to tracking, based on the average absolute tracking they experience across the 5 years following grade 3. Students exposed to more tracking experience higher test score growth at almost all points of the distribution in both time frames. Of course, these patterns do not necessarily reflect causal relationships, since test score growth could be impacted by a variety of factors – school and non-school – that are correlated with tracking.

For a more rigorous examination that allows us to control for potential confounders, we use regression analysis to examine how tracking affects test score mobility for students near the top and bottom of the initial test score distribution. As our dependent variables, we use future percentile ranks (defined on a scale from 0 to 1) for students who started at the 25th and 75th percentiles of the initial test score distribution. Since the data can be sparse at the school-cohort level, we use a parametric approach to estimate these variables. For each school-cohort, we

[34] These patterns are documented in Appendix Figure C8.
[35] This is also similar to the income mobility literature (e.g., Chetty et al. 2014; Hashim et al. 2020), which relates children's positions in the income or education distributions to those of their parents.

regress students' *t*-years-later test score percentiles on their grade 3 percentiles separately for students initially above and below the statewide median. We use the estimated coefficients to predict the position *t* years after grade 3 for students at the 25th and 75th percentiles *p* for cohort *c* from school *s*: $\hat{Y}_{pcs}^{3+t}$.

With these variables in hand, we estimate separately for the 25th and 75th percentiles the following school-cohort level regressions for the various time horizons, weighted by the number of students in each school-cohort:

$$\hat{Y}_{pcs}^{3+t} = \beta_0 + \beta_1 T_{cs}^{4-5} + \beta_2 T_{cs}^{6-8} + X_{cs}\Gamma + \alpha_s + \delta_c + \epsilon_{pcs},$$

where T_{cs}^{4-5} and T_{cs}^{6-8} are school-cohort exposure to (absolute or relative) tracking in grades 4-5 and 6-8.[36] Dividing tracking exposure by grade level enables us to examine whether there are different effects for early versus later exposure to tracking, as well as to conduct placebo tests for whether future tracking is correlated with current outcomes. To control for the initial ability distribution and subsequent attrition of the school-cohort, the vector X_{cs} includes the mean and standard deviation of grade 3 test scores, as well as the 10th, 25th, 75th, and 90th percentiles, and the fractions enrolled each year after grade 3. All regressions also include school and cohort fixed effects. Thus, the coefficients on tracking exposure are identified from variation across cohorts within schools over time, conditional on initial levels of and heterogeneity in ability and persistence in the system.

Because tracking increases in middle school, students with low test scores may experience less tracking simply because they are retained and spend more time in earlier grades. In addition, parents may change schools within district in response to the interaction between a school's tracking policy and their child's ability level. To overcome these endogeneity issues, we instrument the school-cohort's actual tracking exposure with the district-level tracking exposure among the subset of students in the same cohort who advance one grade each year. This additionally helps address measurement error. A lingering concern with interpreting our estimates as the causal impact of greater tracking (and the associated bundle of policies and practices that comes along with increased sorting by ability across classes) on test score growth is that changes in district-level tracking may coincide with unrelated locality or policy changes

[36] We take the mean of the relevant tracking measure over students in the school-cohort in the given year since grade 3, where some students may be in different schools or grades. Then, we take the simple average of these school-cohort-year means across the relevant years since grade 3 as defined by students who progress normally.

that impact student achievement.

Table 7 presents our ordinary least-squares (OLS) and instrumental variables (IV) results for students at the 25th and 75th percentiles. The first 4 columns are based on our absolute measure of tracking, while the second 4 are for relative tracking. Within the columns, each pair of elementary and middle school tracking exposure coefficient estimates is from a separate regression. Moving down the rows, the number of years since grade 3 increases from 2 to 5, as the typical student progresses from grade 5 to 8.

For initially lower-achieving students (columns 1-2 and 5-6), we find that exposure to elementary and middle school tracking generally has little effect on predicted performance 2, 3, 4, and 5 years after grade 3. The point estimates tend to be negative for elementary school tracking exposure and positive for middle school tracking exposure, though most are statistically insignificant. Exceptions are that, in the OLS specifications, exposure to elementary school tracking has a statistically significant negative impact on predicted achievement 3 years after grade 3, and exposure to middle school tracking has a positive impact on predicted performance 5 years after grade 3. For 5 years out, the point estimates from IV are very similar to those for OLS but carry much larger standard errors and are far from statistically significant. Nonetheless, quantifying the 5-year-later estimate from IV for reference, the point estimate of 0.029 for absolute tracking implies that a 1 standard deviation (or 0.10) increase in exposure to middle school tracking is associated with a minimal (0.3 percentile) increase in achievement rank.

Among initially higher-achieving students (columns 3-4 and 7-8), we see more consistent evidence of benefits associated with exposure to tracking in middle school, though again little impact of exposure in elementary school. Reassuringly, we do not see any impact of future exposure to middle-school tracking 2 years after grade 3, when most students are still in elementary school. For absolute tracking and 5 years after grade 3, the magnitude of the IV estimate is double that for the students at the 25th percentile, though still quite small. For students initially at the 75th percentile, a 1 standard deviation increase in middle-school tracking maps to a 0.6 percentile increase in rank. The impacts are larger 4 years after grade 3 at the top (i.e., close to 1 percentile), which is a more accurate horizon to consider for test scores for high achievers since their percentile ranks are almost always concurrent rather than imputed from the prior year.

Our finding of small effects of tracking on student performance is in contrast with some recent work finding relatively large benefits at both ends of the ability distribution. For example,

Card and Giuliano (2016) find that high-performing non-gifted students tracked in classrooms with gifted students experience achievement gains on the order of 0.5 standard deviations and, similarly, Cohodes (2020) finds that the marginal high-achieving students admitted to classes with accelerated coursework experience substantial gains in attainment. On the other end of the achievement distribution, Ballis and Heath (2019) document that students rationed out of being classified as disabled, that are presumably less likely to be downwardly tracked, have worse long-run academic outcomes. Our findings are not necessarily inconsistent with these studies, however, since we are estimating the average effects of more-tracked regimes on all students rather than within-regime effects on marginal students who are tracked upward or downward.

5.2 Distribution of Educational Inputs

We next investigate how math curriculum, class size, and peer quality vary for students at different points of the initial achievement distribution in more- versus less-tracked regimes. This builds on our earlier findings that these factors are correlated with tracking at the school-cohort level.[37] We apply the same two-step estimation strategy specified above, replacing a student's math test score percentile *t* years after grade 3 with alternative outcomes. From the first-stage regressions for each school-cohort, we generate the predicted likelihood of being above or below 8th-grade level math 5 years after grade 3, and average class size and classroom peer quality across grades 4-5 and 6-8 (i.e., 1-2 and 3-5 years after grade 3), for students initially at the 25th and 75th percentiles of the statewide test score distribution. We then relate these predicted variables to the level of tracking exposure in elementary and middle school.

Table 8 displays our results. The layout of the columns matches that of Table 7, with the results shown for different outcomes moving down the rows. Though both OLS and IV results are shown, we streamline the discussion by highlighting the patterns revealed by the IV results.

When we examine curriculum, we see that exposure to tracking is associated with a lower likelihood of being above grade level in math for students initially at the 25th percentile, but a higher likelihood for students at the 75th percentile. Tracking in the elementary grades plays a role in screening lower-achieving students out of advanced coursework, while it is only tracking in the middle school grades that appears to open opportunities for higher-achieving students. For

[37] Unfortunately, we only have access to average teacher experience by school and year, so cannot explore variation across students assigned to different classes.

both lower- and higher-achieving students, we do not find a relationship between the likelihood of being below grade-level math (which embeds grade retention) and exposure to tracking at either grade level.

With respect to class size, we find that for students at both the 25th and 75th percentiles, exposure to absolute tracking in elementary school is associated with smaller average class sizes in elementary school, and exposure to absolute tracking in middle school is associated with smaller average class sizes in middle school. While relative and absolute tracking have quite similar effects on the other inputs, the relationships between relative tracking – which conditions on the class size distribution – and class sizes are more muted. To the extent that tracking is accompanied by smaller class sizes, tracking might benefit both lower- and higher-achieving students. The relationship between tracking and class size is substantially stronger for students at the 25th percentile, which could indicate that high-tracking regimes invest more resources in lower-achieving students, sensitive to the potential for tracking to increase achievement gaps.

The final inputs we consider are the level and standard deviation of classroom peer initial math z-scores. Exposure to tracking widens the gap in peer quality across low- and high-achieving students. For an increase of 0.10 in absolute tracking in the elementary grades, grade 4-5 peer test scores fall by 0.06 and rise by 0.08 standard deviations for students initially at the 25th and 75th percentiles, respectively. The comparable numbers for middle school tracking and grade 6-8 peer test scores are 0.03 and 0.06.[38] Interestingly, exposure to elementary school tracking is associated with lower peer achievement in middle school for low achievers, suggesting that elementary school tracking has persistent negative effects. That tracking tends to lower peer achievement for lower-achieving students and increase peer achievement for higher-achieving students is expected. The final rows in Table 8 show that both types of students experience reductions in heterogeneity in ability in the classroom as students are more sorted by ability across classes, though the magnitudes are greater for students initially at the 25th percentile.[39]

[38] The effects on peer quality are magnified (by up to one-third) if prior-year peer test scores are used in place of initial peer scores.

[39] Conditional on elementary school tracking, we find small positive associations between the standard deviation of peers' initial scores in elementary school and exposure to future tracking in middle school for both lower- and higher-achieving students that is puzzling. Though one possibility is that the degree of middle school tracking could be correlated with where in the achievement distribution tracking is happening during elementary school, we find the same positive association at the 50th percentile.

Taken together, these results suggest that, for higher-achieving students, the positive association between middle school tracking and test score mobility noted in the previous section may operate through exposure to higher quality peers, smaller and more homogenous classes, and access to more advanced coursework. That tracking does not harm lower-achieving students, despite exposure to lower quality peers, may arise from its association with smaller class sizes and more tailored curriculum. It is also possible that self-perception is improved when relative rank among classroom peers is higher (e.g., Malamud et al., 2023; Murphy and Weinhardt, 2020).

6. Conclusion

Very little is known about the nature and scope of ability tracking in the US. In this paper, we use detailed administrative data from Texas to create measures of within-school tracking for grades 4 through 8 for almost every public school in Texas for the 2010-11 to 2018-19 school years. Our data-driven approach allows us to capture both formal and informal tracking within schools, enabling us to provide a comprehensive picture of tracking, including: how sorted students are by ability across classrooms within schools across grades, how schools operationalize tracking, which schools are more likely to track, and how tracking is related to student performance.

We find tracking is prevalent and that there is a great deal of variation in the level of tracking across districts. In addition, in contrast to the popular perception, we find that the amount of ability tracking that takes place within schools is far greater than the amount of ability sorting that occurs across districts and schools. Within-school sorting based on prior test scores is also far greater than within-school sorting based on race/ethnicity and SES. Further, while within-school ability tracking increases substantially as students move from elementary to middle school, there is no such increase in sorting by race/ethnicity and SES.

Tracking is highly coordinated across school-grades within districts, likely reflecting common policies and practices. On the ground, it appears to be operationalized through more aggressive classification of students in special needs categories, such as gifted or disabled, and increased curricular differentiation. Among the most important predictors of tracking is heterogeneity in the prior achievement of students within a school-grade cohort and the type of school, with charter schools tracking less than traditional public schools. Though the results are

imprecise, we do not find that the Democratic share of county voters matters.

Finally, when we examine the implications of tracking for achievement gains, we find that exposure to tracking in elementary and middle school largely holds low-achieving students harmless while providing small benefits to initially high-achieving students. Students in tracked regimes are served in smaller and more homogeneous classes and have access to more differentiated coursework. All in all, our findings suggest that tracking and the typical associated bundle of instructional practices do not harm low-achieving students on average but weakly increase inequities in educational outcomes.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Alzen, J., & Domingue, B. (2013). A characterization of sorting and implications for value-added estimates. *Online Submission.* http://eric.ed.gov/?id=ED545383

Bacher-Hicks, A., & Avery, C. (2018). Panel paper: The effect of classroom assignment policies on equitable access to high-quality teachers.

Ballis, B., & Heath, K. (2021). The long-run impacts of special education. *American Economic Journal: Economic Policy*, *13*(4), 72-111.

Bauer, P., & Riphahn, R. (2006). Timing of school tracking as a determinant of intergenerational transmission of education. *Economics Letters*, *91*(1), 90-7.

Berends, M., & Donaldson, K. (2016). Does the organization of instruction differ in charter schools? Ability grouping and students' mathematics gains. *Teachers College Record*, 118(11).

Betts, J. R. (2011). The economics of tracking in education. *Handbook of the Economics of Education*, *3*(1), 341-81.

Betts, J. R., & Shkolnik, J. L. (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review*, *19*(1), 1-15.

Card, D., & Giuliano, L. (2016). Can tracking raise the test scores of high-ability minority students? *American Economic Review*, *106*(10), 2783-2816.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1553-1623.

Clark, D., & Del Bono, E. (2016). The long-run effects of attending an elite school: evidence from the United Kingdom. *American Economic Journal: Applied Economics*, *8*(1), 150-76.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778-820.

Clotfelter, C. T., Hemelt, S. W., Ladd, H. F., & Turaeva, M. (2021). *School segregation in the era of color-blind jurisprudence and school choice*. (EdWorkingPaper: 21-101). Retrieved 6.16.21 from Annenberg Institute at Brown University: https://doi.org/10.26300/wc3k-ht80.

Cohodes, S. R. (2020). The long-run impacts of specialized programming for high-achieving students. *American Economic Journal: Economic Policy*, *12*(1), 127-66.

Collins, C. A., & Gan, L. (2013). Does sorting students improve scores? An analysis of class composition. National Bureau of Economic Research Working Paper No. 18848.

Cortes, K. E., & Goodman, J.S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of double-dose algebra on student achievement. *American Economic Review*, *104*(5), 400-5.

Dalane, K., & Marcotte, D. E. (2020). *The segregation of students by income in public schools*. (EdWorkingPaper: 20-338). Retrieved 6.26.21 from Annenberg Institute at Brown University: https://doi.org/10.26300/kqkr-0c04.

De Brey, C., Snyder, T.D., Zhang, A., & Dillow, S.A. (2021). *Digest of Education Statistics 2019* (NCES 2021-009). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Table 101.40, Retrieved 6.4.2021 from https://nces.ed.gov/programs/digest/d19/tables/dt19_101.40.asp?current=yes.

Dieterle, S., Guarino, C.M., Reckase, M.D., & Wooldridge, J. M. (2014). How do principals assign students to teachers? Finding evidence in administrative data and the implications for value added. *Journal of Policy Analysis and Management*, *34*(1), 32-58.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), 1739-74.

Dustmann, C., Puhani, P. A., & Schönberg, U. (2017). The long-term effects of early track choice. *The Economic Journal*, *127*(603), 1348–80.

Epple, D., Newlon, E., & Romano, R. (2002). Ability tracking, school competition, and the distribution of educational benefits. *Journal of Public Economics*, *83*(1), 1-48.

Figlio, D. N., & Page, M. E. (2002). School choice and the distributional effects of ability tracking: Does separation increase inequality? *Journal of Urban Economics*, *51*(3), 497-514.

Fu, C., & Mehta, N. (2018). Ability tracking, school and parental effort, and student achievement: a structural model and estimation. *Journal of Labor Economics*, *36*(4), 923-79.

Goodman, J. (2019). The labor of division: Returns to compulsory high school math coursework. *Journal of Labor Economics* 37(4), 1141-82.

Hanushek, E.A., & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, *116*(510), C63–76.

Hashim, S. A., Kane, T.J., Kelley-Kemple, T., Laski, M.E., & Staiger, D.O. 2020. Have income-based achievement gaps widened or narrowed? National Bureau of Economic Research Working Paper No. 27714.

Hellerstein, J. K., McInerney, M., & Neumark, D. (2011). Neighbors and coworkers: the importance of residential labor market networks. *Journal of Labor Economics*, *29*(4), 659-95.

Horvath, H. (2015). Classroom assignment policies and implications for teacher value-added estimation. Working Paper.

Kalogrides, D., & Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within school. *Educational Researcher,* 42(6), 304-16.

Loveless, T. (2013). *The 2013 Brown Center report on American education: How well are American students learning*. Brookings Institute.

Lefgren, L. (2004). Educational peer effects and the Chicago public schools. *Journal of Urban Economics*, *56*(2), 169-91.

Macartney, H., & Singleton, J. D. (2018). School boards and student segregation. *Journal of Public Economics* 164, 165-82.

Malamud, O., Mitrut, A.,Pop-Eleches, C., & Urquiola, M. S., (2023). Self-, peer-, and teacher perceptions under school tracking. SSRN Working Paper No. 4408041.

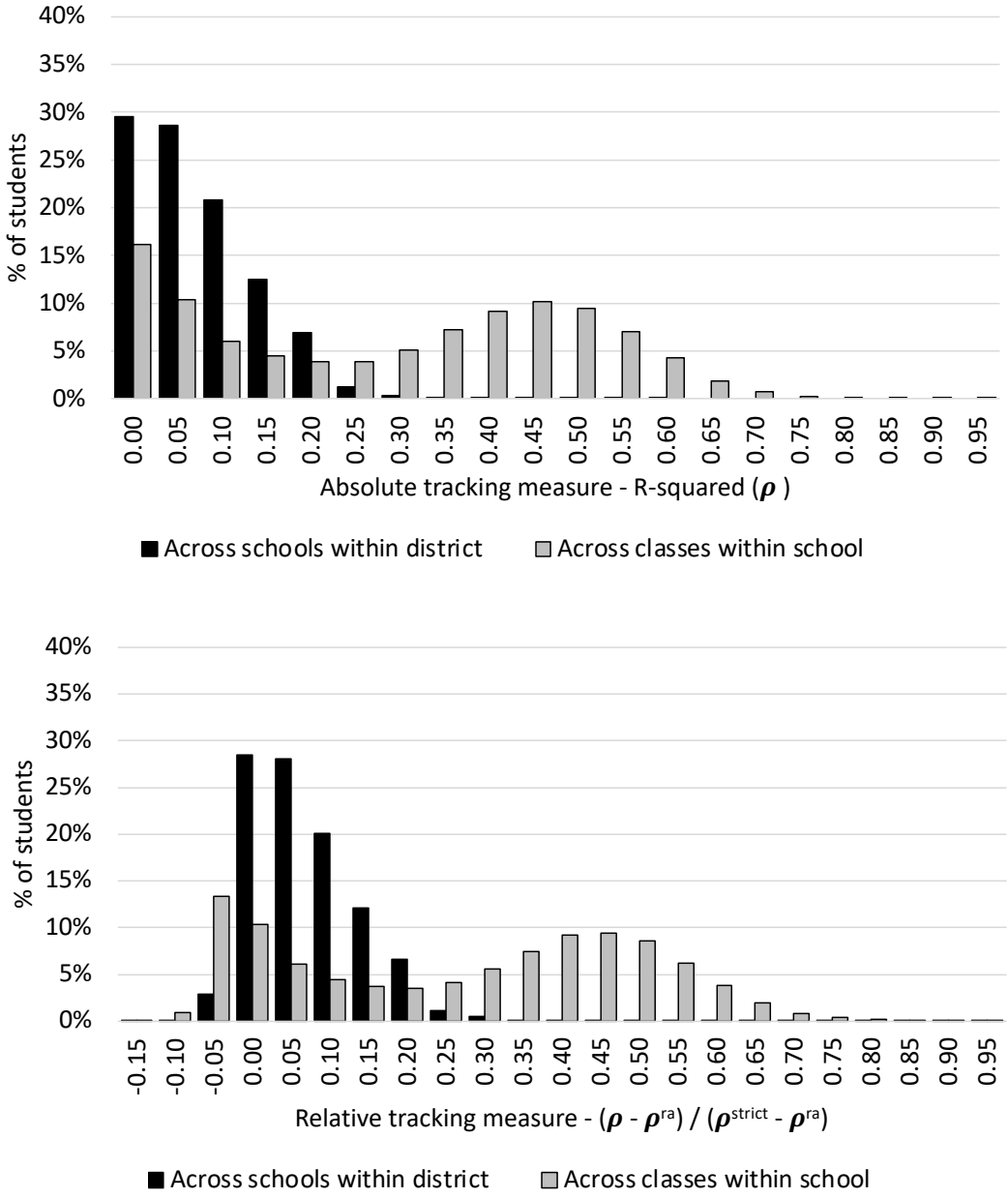
Murphy, R. & Weinhardt, F. (2020). Top of the class: the importance of ordinal rank. *The Review of Economic Studies* 87(6), 2777–2826.

Reardon, S. F. (2019). Educational opportunity in early and middle childhood: Using full population administrative data to study variation by place and age. *The Russell Sage Foundation Journal of the Social Sciences*, 5(2), 40-68.

Rees, D. I., Brewer, D. J., & Argys, L. M. (2000). How should we measure the effect of ability grouping on student performance? *Economics of Education Review*. 19 (1), 17–20.

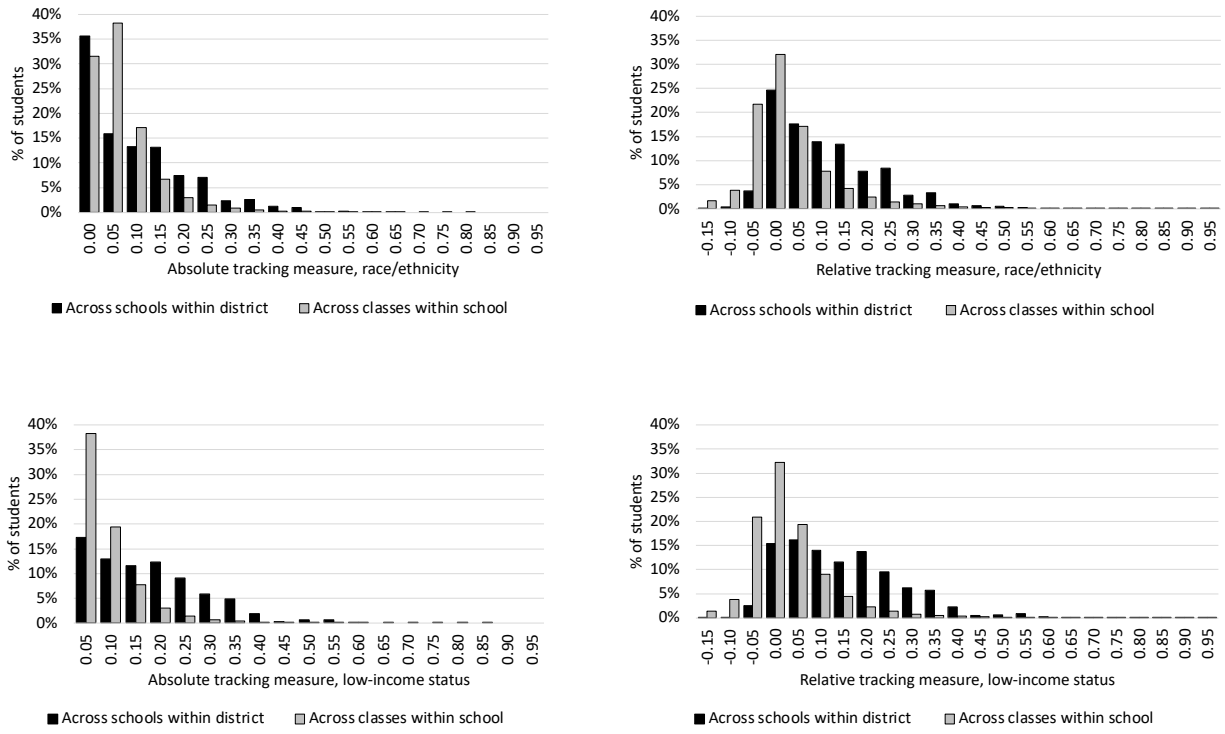
Texas Education Agency. (2020). *Comprehensive biennial report on Texas public schools*. Austin, TX. December, p.179, Table 7.2. https://tea.texas.gov/reports-and-data/school-performance/accountability-research/comprehensive-report-on-texas-public-schools.

Figure 1. Distribution of Tracking Within and Across Schools



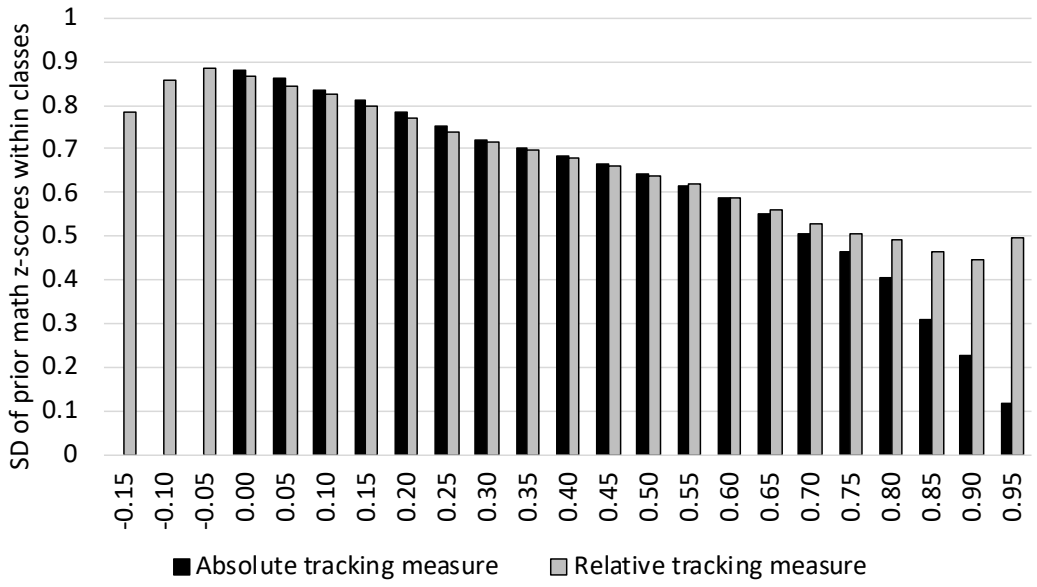
Notes: The top and bottom panels show the student-weighted distributions of the absolute and relative tracking measures, respectively. The grey bars show the distributions when tracking is defined to be across classes within a school-grade-year. For comparison, the black bars show the distributions when tracking is defined to be across schools within a district-grade-year. In all cases, the samples include only district-grade-year cells with more than one school. For this exercise, charter schools are assigned to the geographic districts within which they are located, rather than their administrative districts.

Figure 2. Distribution of Tracking, by Race/Ethnicity and Low-Income Status



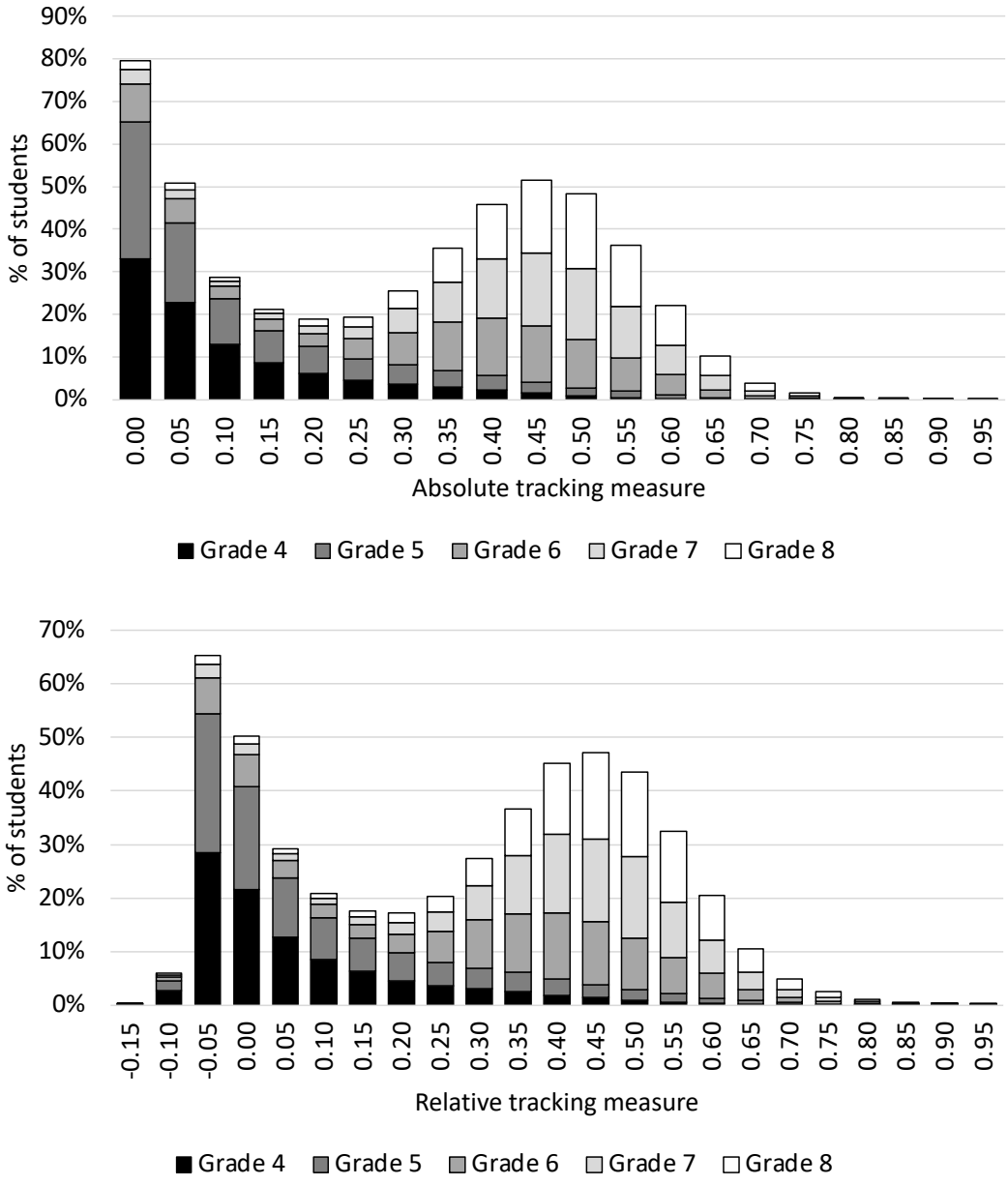
Notes: The panels show the student-weighted distributions of the absolute and relative tracking measures based on student race/ethnicity (defined as Black or Hispanic vs. non-Black and non-Hispanic) in the top panels and by low-income status in the bottom panels. For other details, see the notes to Figure 1.

Figure 3. Within-Class Standard Deviation of Prior Scores, by Level of Tracking



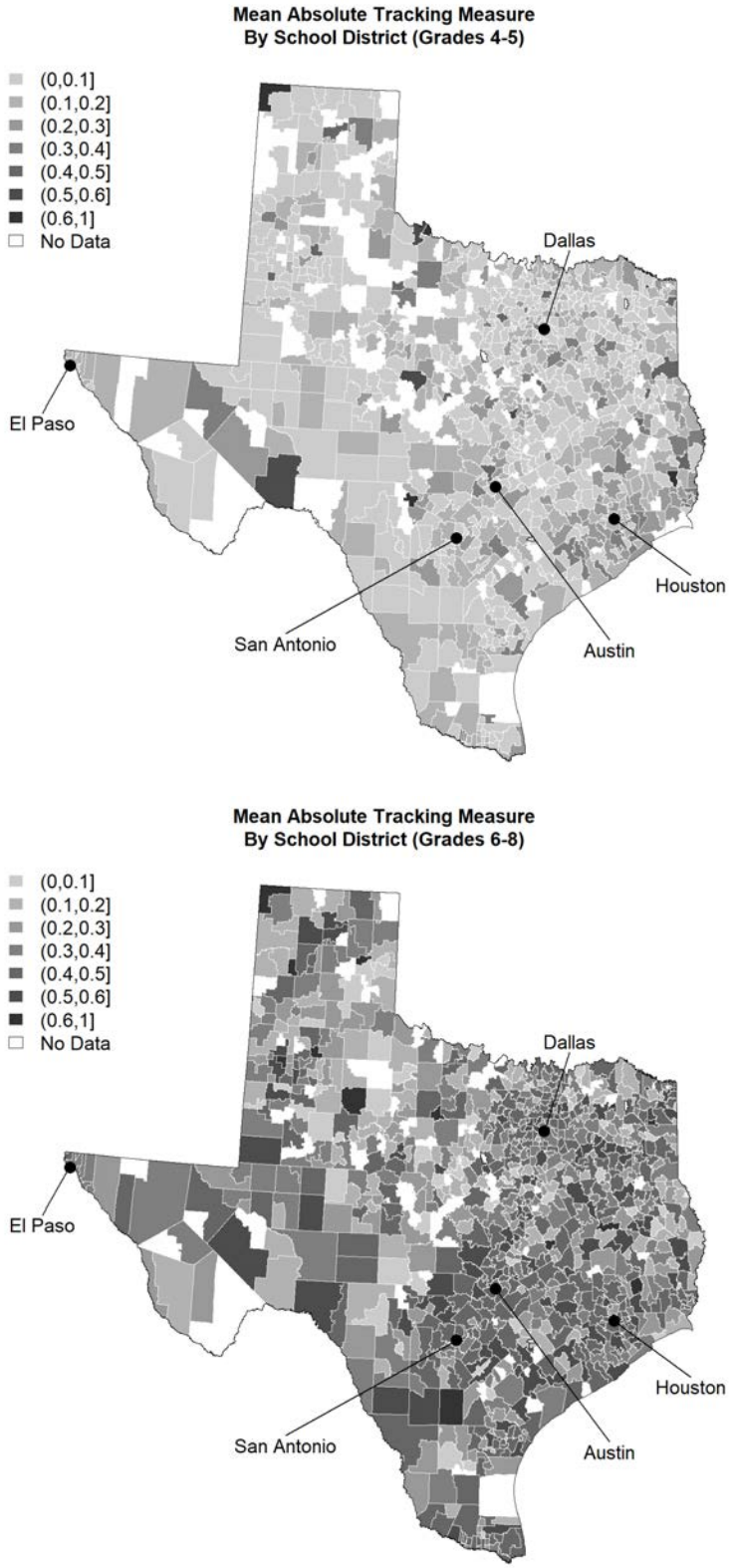
Notes: This figure shows the (student-weighted) distribution of within-class standard deviations of students' prior math z-scores, broken down by levels of absolute and relative tracking.

Figure 4. Extent of Tracking by Grade



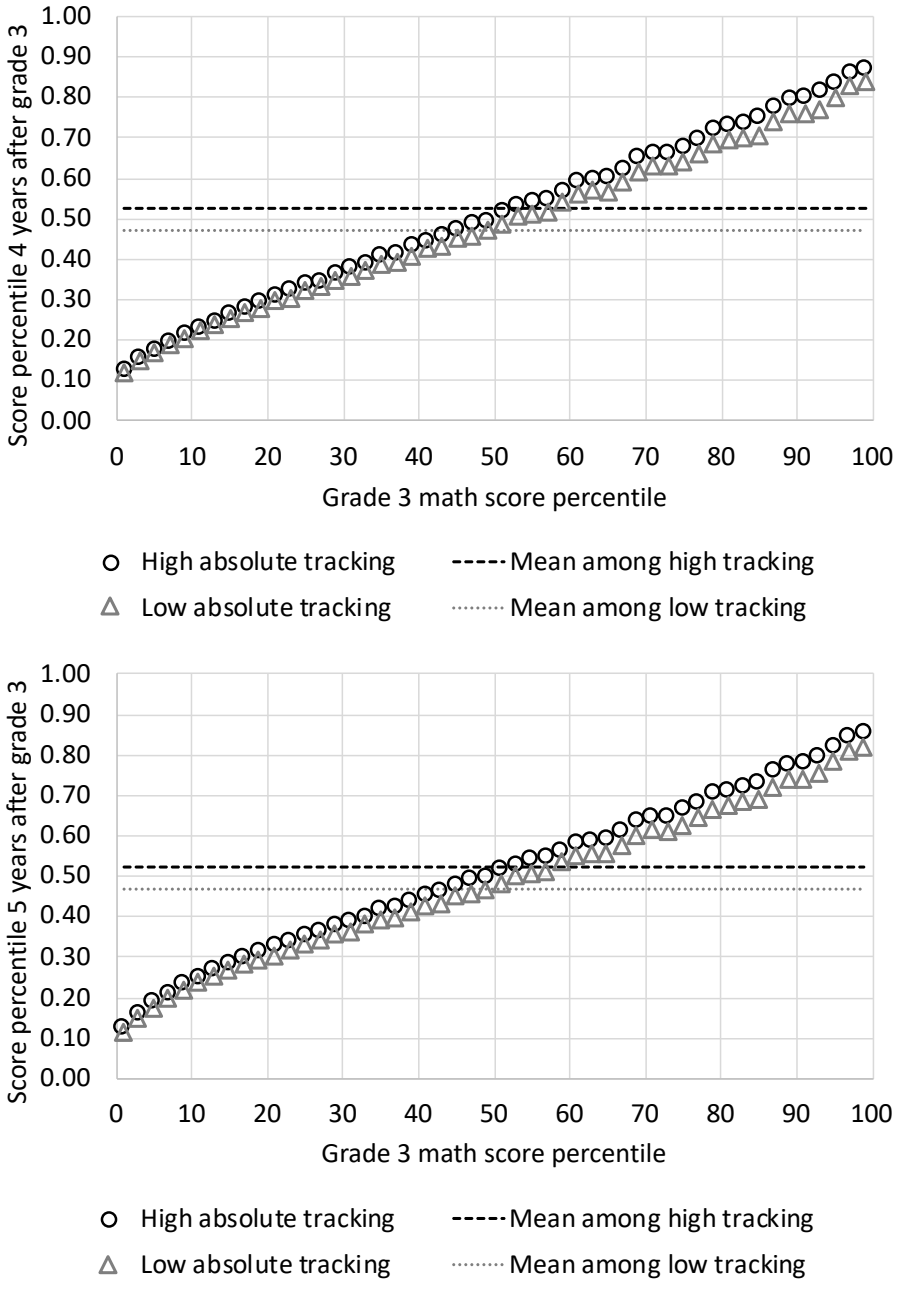
Notes: The top (bottom) panel shows the student-weighted distribution of the absolute (relative) tracking measure by grade, for grades 4-8.

Figure 5. Tracking across Districts in Texas



Notes: The maps show geographic variation across districts in average (student-weighted) school-grade-year absolute tracking for grades 4-5 (top panel) and grades 6-8 (bottom panel).

Figure 6. Achievement Mobility and Tracking



Notes: The top (bottom) panel shows the average math score percentile 4 (5) years after grade 3 by percentile in the 3rd grade math test score distribution, separately for students in school grade 3 cohorts with above vs. below median exposure to absolute tracking. To measure exposure, we first calculate the student-weighted average of absolute tracking over a school-cohort in each year since grade 3, and then take the simple average across years 2 through 5.

Table 1. Summary Statistics

Variable	All grades	Grade				
		4	5	6	7	8
Tracking:						
Absolute tracking measure	0.318	0.133	0.162	0.376	0.450	0.469
	(0.213)	(0.131)	(0.161)	(0.183)	(0.155)	(0.152)
Relative tracking measure	0.297	0.098	0.136	0.358	0.435	0.462
	(0.236)	(0.156)	(0.188)	(0.205)	(0.177)	(0.173)
Fraction of students:						
With identifiable math course	0.931	0.959	0.955	0.926	0.915	0.901
	(0.130)	(0.148)	(0.141)	(0.110)	(0.108)	(0.129)
Missing prior test scores	0.062	0.061	0.062	0.058	0.061	0.069
	(0.042)	(0.041)	(0.042)	(0.036)	(0.038)	(0.051)
Male	0.513	0.513	0.513	0.513	0.512	0.511
	(0.050)	(0.050)	(0.050)	(0.052)	(0.049)	(0.049)
White	0.293	0.284	0.287	0.295	0.297	0.301
	(0.257)	(0.261)	(0.261)	(0.254)	(0.254)	(0.256)
Hispanic	0.513	0.521	0.519	0.511	0.508	0.506
	(0.293)	(0.301)	(0.299)	(0.287)	(0.286)	(0.288)
Black	0.130	0.130	0.130	0.130	0.130	0.130
	(0.155)	(0.164)	(0.162)	(0.150)	(0.148)	(0.149)
Asian	0.039	0.039	0.039	0.040	0.039	0.039
	(0.075)	(0.080)	(0.078)	(0.073)	(0.071)	(0.070)
Other race/ethnicity	0.025	0.027	0.026	0.025	0.025	0.024
	(0.023)	(0.026)	(0.025)	(0.022)	(0.021)	(0.020)
Low income	0.613	0.637	0.630	0.612	0.600	0.588
	(0.269)	(0.279)	(0.278)	(0.264)	(0.260)	(0.259)
Limited English proficient	0.164	0.231	0.195	0.154	0.130	0.110
	(0.169)	(0.210)	(0.190)	(0.150)	(0.132)	(0.117)
LEP self-contained	0.096	0.179	0.149	0.067	0.045	0.037
	(0.156)	(0.198)	(0.178)	(0.125)	(0.102)	(0.088)
Disability - Physical	0.013	0.012	0.013	0.013	0.012	0.012
	(0.012)	(0.015)	(0.015)	(0.011)	(0.010)	(0.010)
Disability - Non-physical	0.083	0.081	0.085	0.086	0.083	0.081
	(0.033)	(0.037)	(0.037)	(0.032)	(0.030)	(0.030)
Disability - Restricted setting	0.016	0.016	0.016	0.016	0.016	0.015
	(0.015)	(0.019)	(0.018)	(0.014)	(0.012)	(0.012)
Gifted	0.100	0.094	0.104	0.102	0.102	0.100
	(0.085)	(0.084)	(0.087)	(0.088)	(0.083)	(0.081)
Curricular differentiation	0.092	0.005	0.009	0.029	0.076	0.340
	(0.169)	(0.033)	(0.054)	(0.097)	(0.146)	(0.165)
Average class size	19	17	20	19	19	19
	(5)	(4)	(5)	(5)	(5)	(6)
Average teacher experience	10.745	11.084	11.095	10.603	10.445	10.498
	(2.870)	(2.963)	(2.982)	(2.900)	(2.737)	(2.687)
Number of school-grade-years	115,792	34,725	32,197	17,701	15,442	15,727
Number of schools	6,695	4,532	4,390	2,737	2,154	2,162
Number of districts	1,128	1,008	1,016	1,051	1,043	1,067

Notes: The sample is students in regular instructional public schools over the period 2011 to 2019 (school years 2010-11 to 2018-19), among school-grade-years with at least two separate math classes. Each column shows the means and standard deviations for students in the grade indicated in the column heading for the variables indicated by the row headings. Low-income students are those who are eligible for free or reduced-price meals or certain public assistance programs (such as TANF). Limited English proficient students can be served in bilingual or English as a Second Language (ESL) programs. These programs can be structured to integrate students who are proficient in English for instruction in the core subjects, such as through bilingual two-way immersion or ESL pullout, or self-contained, such as bilingual non-two-way and ESL content-based programs. Physical disabilities include disabilities such as orthopedic impairment, auditory or visual impairment, and traumatic brain injury, while most other disabilities are emotional and learning disabilities. We classify special education instructional settings as restricted if the student spends less than half of the school day in general education classrooms. Curricular differentiation is measured as one minus the Herfindahl index of concentration, calculated based on the shares of students served under different math course titles.

Table 2. Total Variation in Scores and Characteristics Accounted for by District/School/Class

	Variance in prior test scores accounted for by:			Variance in race/ethnicity accounted for by:			Variance in low-income status accounted for by:		
	District	School	Class	District	School	Class	District	School	Class
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
All students									
All districts	0.09	0.17	0.44	0.26	0.34	0.40	0.19	0.30	0.37
Districts with (minimum) 1 school	0.13	0.14	0.41	0.31	0.32	0.38	0.18	0.20	0.28
Districts with 2-5 schools	0.11	0.17	0.46	0.32	0.37	0.44	0.26	0.34	0.41
Districts with 6+ schools	0.07	0.18	0.44	0.16	0.29	0.36	0.16	0.33	0.39
Grades 4-5									
All districts	0.07	0.16	0.29	0.27	0.37	0.42	0.20	0.34	0.39
Districts with (minimum) 1 school	0.11	0.14	0.26	0.31	0.33	0.39	0.18	0.22	0.28
Districts with 2-5 schools	0.08	0.16	0.30	0.31	0.39	0.45	0.25	0.36	0.42
Districts with 6+ schools	0.05	0.17	0.29	0.17	0.34	0.40	0.17	0.39	0.44
Grades 6-8									
All districts	0.10	0.17	0.54	0.26	0.32	0.38	0.19	0.28	0.35
Districts with (minimum) 1 school	0.13	0.14	0.50	0.31	0.32	0.38	0.18	0.19	0.28
Districts with 2-5 schools	0.13	0.17	0.57	0.32	0.36	0.43	0.26	0.32	0.40
Districts with 6+ schools	0.07	0.18	0.54	0.14	0.26	0.33	0.15	0.30	0.37

Notes: Districts are grouped by the minimum number of schools for any grade-year across grades 4-8 and years 2011-2019. Charter schools are assigned to their geographic districts rather than their administrative districts. The R-squared reported in each cell is from a regression of the variable indicated in the column header (i.e., prior-year math test z-scores, an indicator for Black or Hispanic, or an indicator for low-income status) on a set of indicators for each district, school, or class, as indicated in the column sub-header.

Table 3. Correlations Between Tracking by Prior Math Scores across Subjects

			Math		ELA		Science		Social Studies	
			Absolute	Relative	Absolute	Relative	Absolute	Relative	Absolute	Relative
All Grades	Math	Absolute	1							
		Relative	0.99	1						
	ELA	Absolute	0.83	0.83	1					
		Relative	0.80	0.81	0.99	1				
	Science	Absolute	0.76	0.76	0.85	0.83	1			
		Relative	0.74	0.75	0.83	0.84	0.99	1		
	Social Studies	Absolute	0.71	0.71	0.83	0.83	0.90	0.89	1	
		Relative	0.69	0.70	0.82	0.83	0.88	0.89	0.99	1
Grades 4-5	Math	Absolute	1							
		Relative	0.99	1						
	ELA	Absolute	0.90	0.89	1					
		Relative	0.89	0.90	0.99	1				
	Science	Absolute	0.88	0.88	0.94	0.93	1			
		Relative	0.87	0.88	0.93	0.94	0.99	1		
	Social Studies	Absolute	0.86	0.86	0.94	0.93	0.96	0.95	1	
		Relative	0.85	0.86	0.93	0.94	0.95	0.96	0.99	1
Grades 6-8	Math	Absolute	1							
		Relative	0.98	1						
	ELA	Absolute	0.68	0.66	1					
		Relative	0.64	0.65	0.98	1				
	Science	Absolute	0.59	0.57	0.73	0.70	1			
		Relative	0.57	0.57	0.71	0.71	0.99	1		
	Social Studies	Absolute	0.54	0.52	0.72	0.70	0.83	0.81	1	
		Relative	0.52	0.52	0.70	0.71	0.81	0.82	0.98	1

Notes: For each subject, we calculate our absolute and relative tracking measures. The prior-year math z-score is used in all cases, even for calculating tracking in non-math subjects. This table shows, for each subject combination, the degree to which tracking by math scores in one subject is correlated with tracking by math scores in another subject. The correlations are student-weighted.

Table 4. Fraction of Variation in Tracking Explained

		Variance in absolute tracking measure accounted for by:			Variance in relative tracking measure accounted for by:		
	No. campuses	District	Dist-grade	Dist-grade-yr	District	Dist-grade	Dist-grade-yr
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
All students							
All districts	6,676	0.13	0.71	0.83	0.12	0.68	0.81
Districts with (min) 1 school	2,801	0.18	0.71	0.94	0.16	0.67	0.93
Districts with 2-5 schools	1,483	0.15	0.75	0.84	0.14	0.72	0.81
Districts with 6+ schools	2,392	0.06	0.69	0.73	0.06	0.66	0.70
Grades 4-5							
All districts	4,866	0.36	0.42	0.59	0.33	0.39	0.56
Districts with (min) 1 school	1,887	0.37	0.46	0.81	0.34	0.43	0.80
Districts with 2-5 schools	1,121	0.40	0.47	0.60	0.38	0.44	0.57
Districts with 6+ schools	1,858	0.32	0.37	0.41	0.28	0.33	0.37
Grades 6-8							
All districts	3,071	0.33	0.51	0.74	0.29	0.48	0.74
Districts with (min) 1 school	1,560	0.39	0.58	0.94	0.33	0.53	0.94
Districts with 2-5 schools	621	0.42	0.59	0.76	0.38	0.56	0.74
Districts with 6+ schools	890	0.14	0.33	0.46	0.14	0.32	0.45

Notes: Each cell in columns 2-7 contains the R-squared from a separate (student-weighted) regression. The left-hand-side variable is either the absolute tracking measure (columns 2-4) or the relative tracking measure (columns 5-7), calculated within school-grade-year cells and demeaned by grade-year. The right-hand-side variables are a set of group fixed effects, at the level described in the column title. Across the rows, districts are categorized by the minimum number of schools for any grade-year across grades 4-8 and years 2011-2019. Note that charters are assigned to their administrative districts, not the geographic districts within which they reside, since this is the level at which local policies are determined.

Table 5. Tracking Policies, Absolute Measure of Tracking

	[1]	[2]	[3]	[4]	[5]
Fraction of students:					
Limited English proficient	0.047*	0.049**	0.049**	0.060***	0.026
	(0.025)	(0.024)	(0.024)	(0.017)	(0.024)
LEP self-contained	-0.044	-0.058**	-0.057**	-0.077***	-0.042
	(0.029)	(0.029)	(0.029)	(0.020)	(0.029)
Disability - Physical	0.252**	0.091	0.089	0.116*	0.014
	(0.110)	(0.096)	(0.096)	(0.062)	(0.046)
Disability - Non-physical	0.314***	0.171***	0.174***	0.206***	0.088***
	(0.059)	(0.053)	(0.053)	(0.026)	(0.020)
Disability - Restricted setting	0.275***	0.220**	0.222**	0.086*	0.055
	(0.097)	(0.087)	(0.087)	(0.045)	(0.042)
Gifted	0.203***	0.148***	0.146***	0.149***	0.077***
	(0.042)	(0.041)	(0.041)	(0.035)	(0.022)
Curricular differentiation		0.198***	0.197***	0.262***	0.251***
		(0.026)	(0.026)	(0.030)	(0.031)
Average class size		-0.008***	-0.008***	-0.007***	-0.006***
		(0.000)	(0.000)	(0.000)	(0.000)
Average teacher experience		0.005***	0.005***	0.005***	0.001**
		(0.001)	(0.001)	(0.001)	(0.001)
Mean, SD lagged math test scores	Yes	Yes	Yes	Yes	Yes
Cohort test score percentiles	No	No	Yes	Yes	Yes
Fixed effects (x grade)	None	None	None	District	School
R-squared	0.57	0.61	0.61	0.74	0.82
Number of observations	113,239	112,984	112,984	112,873	112,071
Number of clusters	890	890	890	865	865

Notes: The dependent variable is absolute tracking for the cohort (i.e., school-grade-year cell), with observations weighted by cohort enrollment. In addition to the coefficients displayed in the table, all specifications contain the following controls: grade and year indicators, log of cohort enrollment, the mean and standard deviation of cohort prior math test scores, indicators for whether the school has grade 5 and/or grade 7, log of school district total enrollment, log of tax-assessed property value in the district, and indicators for whether the district is classified as suburban, town, or rural (with urban districts the omitted category). Where indicated, the covariates also include percentiles of the distribution of cohort previous scores (i.e., 10th, 25th, 75th, and 90th percentiles) and fixed effects at the district-by-grade or school-by-grade level. Standard errors are clustered by district, and charter schools are assigned to the geographic districts within which they reside. *** p<0.01, ** p<0.05, * p<0.10

Table 6. Determinants of Tracking, Absolute Measure of Tracking

	[1]	[2]	[3]	[4]	[5]	[6]
Mean lagged z-score	0.016**	-0.017***	-0.016	-0.007	-0.017	-0.005
	(0.006)	(0.006)	(0.010)	(0.017)	(0.012)	(0.011)
Std. dev. lagged z-score		0.298***	0.299***	0.200***	0.205***	0.190***
		(0.017)	(0.016)	(0.014)	(0.012)	(0.011)
Magnet school	0.024	0.023	0.023*	0.023*	0.033***	0.014
	(0.016)	(0.014)	(0.014)	(0.014)	(0.010)	(0.015)
Charter school	-0.135***	-0.136***	-0.135***	-0.135***	-0.149***	n/a
	(0.014)	(0.014)	(0.014)	(0.014)	(0.015)	
District private school share			-0.253*	-0.258*	n/a	n/a
			(0.133)	(0.133)		
County Democratic vote share			-0.008	-0.007	n/a	n/a
			(0.032)	(0.032)		
Fraction of students:						
Hispanic			-0.034	-0.035	-0.024	-0.009
			(0.022)	(0.022)	(0.019)	(0.015)
Black			-0.028	-0.029	-0.035*	-0.002
			(0.023)	(0.023)	(0.021)	(0.019)
Asian			0.017	0.014	-0.001	-0.022
			(0.033)	(0.033)	(0.033)	(0.032)
Other race/ethnicity			-0.009	-0.007	-0.078*	-0.052
			(0.082)	(0.082)	(0.040)	(0.036)
Low income			0.025	0.026	0.005	-0.015
			(0.021)	(0.021)	(0.012)	(0.012)
Limited English proficient			0.011	0.010	-0.004	0.012
			(0.019)	(0.019)	(0.011)	(0.016)
Cohort test score percentiles	No	No	No	Yes	Yes	Yes
Fixed effects (x grade)	None	None	None	None	District	School
R-squared	0.56	0.58	0.58	0.58	0.73	0.81
Number of observations	113,167	113,167	113,167	113,167	113,056	112,263
Number of clusters	890	890	890	890	865	865

Notes: The dependent variable is absolute tracking for the cohort (i.e., school-grade-year cell), with observations weighted by cohort enrollment. In addition to the coefficients displayed in the table, all specifications contain the following controls: grade and year indicators, log of cohort enrollment, indicators for whether the school has grade 5 and/or grade 7, log of school district total enrollment, log of tax-assessed property value in the district, and indicators for whether the district is classified as suburban, town, or rural (with urban districts the omitted category). Where indicated, the covariates also include percentiles of the distribution of cohort previous scores (i.e., 10th, 25th, 75th, and 90th percentiles) and fixed effects at the district-by-grade or school-by-grade level. "District private school share" is the 2010-2016 average share of families with children enrolled in private school. "County Democratic vote share" is the average two-party Democratic vote share across the 2000-2016 presidential elections. Standard errors are clustered by district, and charter schools are assigned to the geographic districts within which they reside. *** $p<0.01$, ** $p<0.05$, * $p<0.10$

Table 7. Effects of Tracking on Achievement Mobility

	Absolute tracking measure				Relative tracking measure			
	Predicted math score percentile for students at 25th percentile in grade 3		Predicted math score percentile for students at 75th percentile in grade 3		Predicted math score percentile for students at 25th percentile in grade 3		Predicted math score percentile for students at 75th percentile in grade 3	
	OLS [1]	IV [2]	OLS [3]	IV [4]	OLS [5]	IV [6]	OLS [7]	IV [8]
	Grade 3 + 2				Grade 3 + 2			
Elementary school tracking	-0.012	-0.019	0.005	0.001	-0.011	-0.016	0.004	0.003
	(0.010)	(0.022)	(0.010)	(0.022)	(0.008)	(0.018)	(0.009)	(0.018)
Middle school tracking	0.000	0.003	0.016	0.004	0.001	-0.001	0.010	0.001
	(0.016)	(0.022)	(0.015)	(0.021)	(0.013)	(0.017)	(0.013)	(0.017)
Dependent variable mean	0.315		0.671		0.315		0.671	
	Grade 3 + 3				Grade 3 + 3			
Elementary school tracking	-0.016*	0.001	0.001	0.020	-0.014**	-0.004	-0.000	0.017
	(0.009)	(0.020)	(0.010)	(0.022)	(0.007)	(0.017)	(0.008)	(0.017)
Middle school tracking	-0.023	-0.018	0.051***	0.052*	-0.021	-0.018	0.044***	0.042*
	(0.019)	(0.025)	(0.019)	(0.028)	(0.016)	(0.020)	(0.016)	(0.022)
Dependent variable mean	0.321		0.663		0.321		0.663	
	Grade 3 + 4				Grade 3 + 4			
Elementary school tracking	-0.012	-0.010	0.009	0.005	-0.011	-0.009	0.007	0.004
	(0.009)	(0.017)	(0.008)	(0.021)	(0.007)	(0.014)	(0.007)	(0.017)
Middle school tracking	0.015	0.021	0.084***	0.094***	0.009	0.015	0.066***	0.071***
	(0.014)	(0.020)	(0.015)	(0.023)	(0.012)	(0.017)	(0.012)	(0.019)
Dependent variable mean	0.326		0.658		0.326		0.658	
	Grade 3 + 5				Grade 3 + 5			
Elementary school tracking	-0.005	-0.001	0.018**	0.016	-0.003	0.000	0.016**	0.017
	(0.010)	(0.020)	(0.009)	(0.020)	(0.008)	(0.017)	(0.007)	(0.016)
Middle school tracking	0.026*	0.029	0.064***	0.061***	0.023*	0.027	0.056***	0.057***
	(0.015)	(0.024)	(0.014)	(0.021)	(0.012)	(0.020)	(0.012)	(0.018)
Dependent variable mean	0.336		0.646		0.336		0.646	

Notes: Each pair of estimated coefficients on elementary and middle school tracking comes from a separate school-cohort level regression, where cohorts are defined based on year of enrollment in grade 3 when initial test scores are recorded. The outcome is the predicted math score percentile some number of years after grade 3, for students with grade 3 math scores in the 25th and 75th percentiles of the statewide distribution. This can be predicted as long as there are 2 or more students with non-missing scores in the top and bottom half of the initial achievement distribution. The estimation sample is balanced across panels by excluding school-cohorts for which any of the predictions are missing. Elementary school (grades 4-5) and middle school (grades 6-8) tracking refer to the simple averages (across a school-cohort) of the tracking measures applicable to each student. In the columns labeled IV, we instrument for tracking with the same averages calculated at the district-cohort level rather than the school-cohort level. When calculating these instruments, we restrict attention to students who have enrollment records for each grade 4-8 and who do not repeat any grades during that period, and we assign charter schools to their geographic districts. All regressions include as controls a set of school and cohort-year fixed effects, the mean and standard deviation of grade 3 math scores (and the scores at the 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentiles) in the school-cohort, and the fraction of the school-cohort with enrollment records for each year (2-5) after grade 3. Standard errors are clustered by district. *** p<0.01, ** p<0.05, * p<0.10

Table 8. Effects of Tracking on Educational Inputs

	Absolute tracking measure				Relative tracking measure			
	Predicted outcome for students at 25th percentile		Predicted outcome for students at 75th percentile		Predicted outcome for students at 25th percentile		Predicted outcome for students at 75th percentile	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
	Likelihood over grade-level in math in grade 3+5				Likelihood over grade-level in math in grade 3+5			
Elementary school tracking	-0.033**	-0.063**	0.006	-0.049	-0.025**	-0.054**	0.006	-0.041
	(0.013)	(0.032)	(0.023)	(0.052)	(0.011)	(0.027)	(0.018)	(0.043)
Middle school tracking	-0.138***	-0.122**	0.290***	0.356***	-0.102***	-0.083*	0.281***	0.344***
	(0.042)	(0.053)	(0.072)	(0.095)	(0.035)	(0.042)	(0.059)	(0.076)
Dependent variable mean	0.115		0.484		0.115		0.484	
	Likelihood under grade-level in math in grade 3+5				Likelihood under grade-level in math in grade 3+5			
Elementary school tracking	0.001	0.012	0.001	0.003	-0.001	0.007	0.001	0.003
	(0.006)	(0.012)	(0.002)	(0.004)	(0.005)	(0.010)	(0.001)	(0.003)
Middle school tracking	-0.027***	-0.003	0.003	-0.003	-0.019**	0.004	0.001	-0.003
	(0.010)	(0.014)	(0.004)	(0.006)	(0.008)	(0.012)	(0.003)	(0.005)
Dependent variable mean	0.043		0.006		0.043		0.006	
	Class size in grades 4-5				Class size in grades 4-5			
Elementary school tracking	-6.064***	-6.441***	-3.737***	-3.713**	-2.732***	-3.216**	-0.827*	-0.825
	(0.703)	(1.818)	(0.750)	(1.864)	(0.413)	(1.369)	(0.438)	(1.372)
Middle school tracking	0.490	1.230	0.405	1.341	0.287	0.912	0.230	0.992
	(0.565)	(0.806)	(0.595)	(0.855)	(0.472)	(0.618)	(0.489)	(0.650)
Dependent variable mean	18.833		19.205		18.833		19.205	
	Class size in grades 6-8				Class size in grades 6-8			
Elementary school tracking	-0.292	0.031	0.071	-0.040	-0.228	0.114	0.101	0.073
	(0.243)	(0.594)	(0.257)	(0.541)	(0.194)	(0.486)	(0.207)	(0.446)
Middle school tracking	-4.082***	-3.083***	-1.864***	-1.032*	-2.237***	-1.520***	-0.514	-0.058
	(0.419)	(0.596)	(0.438)	(0.566)	(0.348)	(0.487)	(0.349)	(0.437)
Dependent variable mean	18.882		20.447		18.882		20.447	
	Grade 4-5 peers' mean initial math z-score				Grade 4-5 peers' mean initial math z-score			
Elementary school tracking	-0.664***	-0.615***	0.782***	0.835***	-0.540***	-0.512***	0.646***	0.704***
	(0.018)	(0.026)	(0.025)	(0.039)	(0.015)	(0.023)	(0.021)	(0.032)
Middle school tracking	0.004	0.008	-0.001	0.011	0.002	0.009	0.004	0.009
	(0.014)	(0.017)	(0.014)	(0.018)	(0.012)	(0.014)	(0.012)	(0.015)
Dependent variable mean	-0.119		0.081		-0.119		0.081	
	Grade 6-8 peers' mean initial math z-score				Grade 6-8 peers' mean initial math z-score			
Elementary school tracking	-0.081***	-0.112***	0.042***	-0.012	-0.071***	-0.101***	0.037***	-0.007
	(0.015)	(0.034)	(0.016)	(0.036)	(0.012)	(0.028)	(0.014)	(0.031)
Middle school tracking	-0.310***	-0.296***	0.524***	0.573***	-0.258***	-0.236***	0.449***	0.485***
	(0.025)	(0.030)	(0.030)	(0.034)	(0.021)	(0.024)	(0.026)	(0.029)
Dependent variable mean	-0.287		0.292		-0.287		0.292	
	SD of Grade 4-5 peers' initial math z-scores				SD of Grade 4-5 peers' initial math z-scores			
Elementary school tracking	-0.512***	-0.520***	-0.380***	-0.394***	-0.419***	-0.438***	-0.304***	-0.327***
	(0.010)	(0.017)	(0.011)	(0.020)	(0.009)	(0.015)	(0.009)	(0.017)
Middle school tracking	0.038***	0.043***	0.040***	0.056***	0.031***	0.036***	0.034***	0.048***
	(0.011)	(0.015)	(0.012)	(0.016)	(0.010)	(0.013)	(0.010)	(0.013)
Dependent variable mean	0.866		0.887		0.866		0.887	
	SD of Grade 6-8 peers' initial math z-scores				SD of Grade 6-8 peers' initial math z-scores			
Elementary school tracking	-0.027***	-0.019	-0.022***	-0.018	-0.024***	-0.016	-0.016***	-0.013
	(0.007)	(0.014)	(0.006)	(0.014)	(0.006)	(0.012)	(0.005)	(0.012)
Middle school tracking	-0.216***	-0.200***	-0.168***	-0.148***	-0.184***	-0.167***	-0.140***	-0.120***
	(0.013)	(0.016)	(0.013)	(0.018)	(0.011)	(0.014)	(0.011)	(0.014)
Dependent variable mean	0.755		0.792		0.755		0.792	

Notes: Each pair of estimated coefficients on elementary (grades 4-5) and middle (grades 6-8) school tracking comes from a separate school-cohort level regression. The outcomes in the top two panels are the predicted likelihood of being enrolled in math courses above (e.g., algebra or geometry) or below (e.g., grade 7 math) the level of grade 8 math 5 years after grade 3, for students from the school-cohort at the 25th and 75th percentiles of the statewide test score distribution in grade 3. The outcomes in the bottom six panels are the predicted average class size and mean and standard deviation of classroom peer achievement (based on math z-scores when first observed) 1-2 years (grades 4-5) and 3-5 years (grades 6-8) after grade 3. The estimation sample is balanced across panels by excluding school-cohorts that are missing any of the dependent variables. For other details, see the notes to Table 7. *** $p<0.01$, ** $p<0.05$, * $p<0.10$

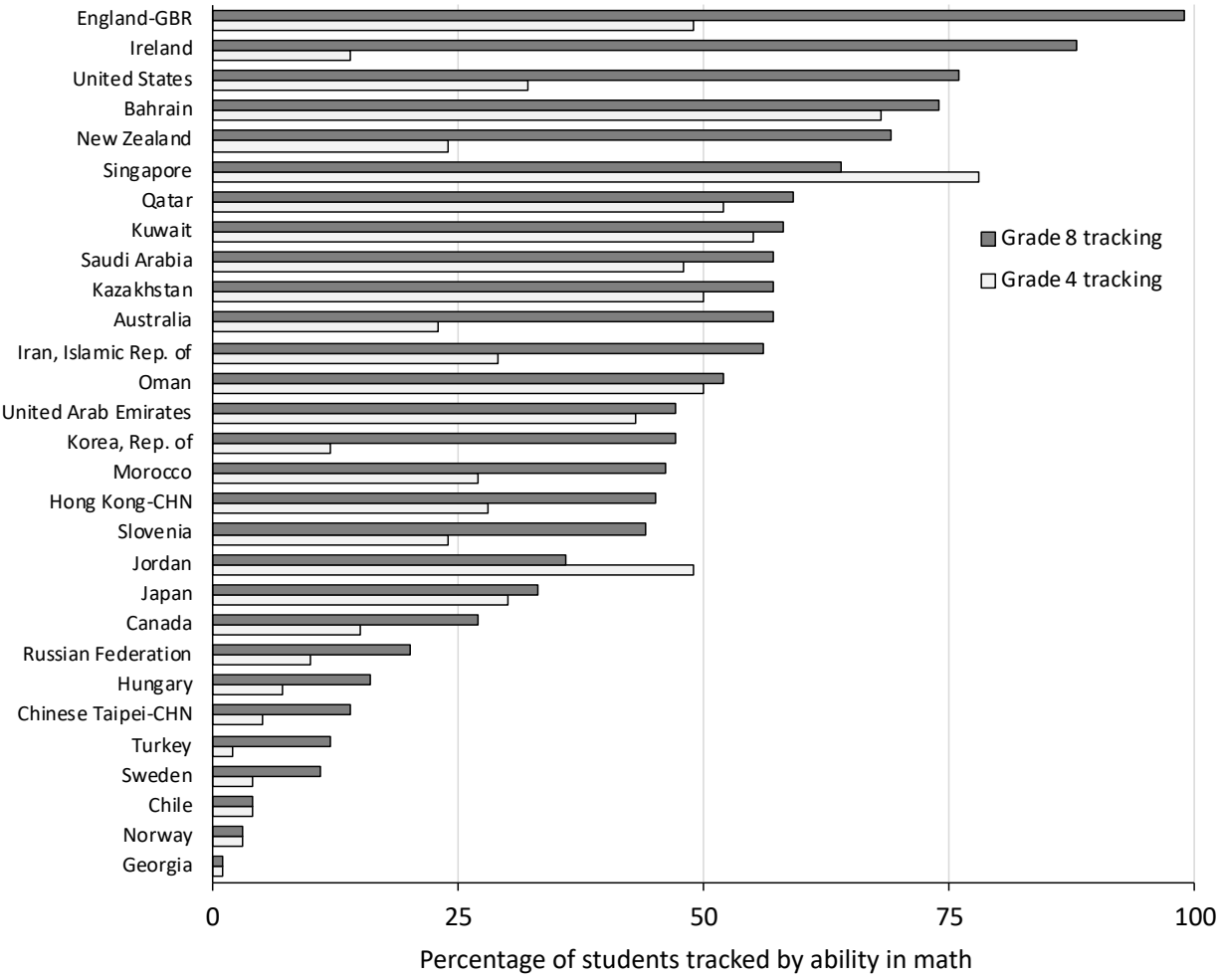
APPENDICES FOR ONLINE PUBLICATION

Appendix A. National and International Survey-Based Patterns in Tracking

School principal survey responses from the National Assessment of Educational Progress (NAEP) reveal that tracking is prevalent in the US. As Table A1 shows, over the past two decades, around one-quarter of 4th graders and three-quarters of 8th graders were in schools that tracked students by ability across classes. These shares have been relatively stable across recent years.

Figure A1 places the US experience in the context of other countries. It reports statistics from the 2015 Trends in International Mathematics and Science Study (TIMSS) for rates of within-school tracking in 4th and 8th grade by participating country. Regardless of the grade, the US exhibits high rates of this form of tracking relative to the typical country surveyed. Few countries exhibit more within-school tracking in 8th grade, with Great Britain and Ireland being among the notable exceptions.

Figure A1. Percentage of Students Tracked by Ability across Math Classes, by Country in 2015



Notes: These statistics are designed to be nationally representative of 2015 student populations and are drawn from TIMSS. The percentages are based on the question "As a general school policy, is student achievement used to assign 4th (8th) grade students to classes for mathematics?" (variables AC6BG10A and BC6BG09A). The percentage shown is the (weighted) share of school administrators responding affirmatively.

Table A1. Percentage of US Students Tracked by Ability across Math Classes

Year	Across-class tracking	
	Grade 4	Grade 8
1990	24	75
1992	—	73
1996	—	71
2000	—	73
2003	—	73
2005	22	73
2007	24	75
2009	28	77
2011	31	76
2013	32	78
2015	32	74
2017	28	—
2019	28	—

Notes: These statistics are drawn from the NAEP Mathematics Assessments and are representative of all US public and nonpublic school students. The percentages shown are based on the (weighted) share of school principals responding affirmatively to the question "Are 4th (8th) graders typically assigned to mathematics classes by ability and/or achievement levels?" (variables C029902, C052001, and C104501 for 4th grade and C028602, C034402, C052901, and C072801 for 8th grade). Note that the wording of the question is different for 4th grade in 2005 and later years since it is phrased as grouping students from different classes by achievement level for math instruction.

Appendix B. Data-Driven Measures of Tracking

The two measures of tracking that we calculate are the "absolute" unadjusted R-squared measure and the "relative" measure that conditions on endogenous constraints on tracking, such as the number of classes and distribution of ability. Both measures are defined at the level of the school-grade-year cell. In this appendix, we provide more details on these measures, their properties, and how they relate to alternative measures.

B.1 Absolute Tracking Measure

Our absolute measure of tracking captures the portion of the variance in prior test scores accounted for by current classes. It is equal to the unadjusted R^2 statistic from a regression of previous test scores on current classroom indicators.

Specifically, let $A = \{a_1, a_2, ...\}$ be the set of students in a school-grade-year cohort, let $C = \{c_1, c_2, ...\}$ be the set of classes, and let b_c be the set of students in class c. Note that $\{b_c\}_{\{c \in C\}}$ is a partition of A, so that every student is in exactly one class. Let x_a be the standardized math test score that student a received at the end of the previous year. Finally, let $N = |A|$ be the number of students, $N_c = |b_c|$ be the size of class c, and $N^C = |C|$ be the number of classes. The cohort mean of prior test scores is $\bar{x} = \frac{1}{N}\sum_{a \in A} x_a$, and the class mean is $\bar{x}_c = \frac{1}{N_c}\sum_{a \in b_c} x_a$.

Given these definitions, the R^2 statistic is:

$$\rho = \frac{\left(\frac{1}{N}\sum_{c \in C} \frac{1}{N_c}\left(\sum_{a \in b_c} x_a\right)^2\right) - \left(\frac{1}{N}\sum_{a \in A} x_a\right)^2}{\left(\frac{1}{N}\sum_{a \in A} x_a^2\right) - \left(\frac{1}{N}\sum_{a \in A} x_a\right)^2} = \frac{\left(\frac{1}{N}\sum_{c \in C} N_c \bar{x}_c{}^2\right) - \bar{x}^2}{\left(\frac{1}{N}\sum_{a \in A} x_a^2\right) - \bar{x}^2}$$

This can be expressed as:

$\rho = \frac{\kappa - \lambda}{\eta - \lambda}$, where $\eta = \frac{1}{N}\sum_{a \in A} x_a^2$, $\kappa = \frac{1}{N}\sum_{c \in C} N_c \bar{x}_c{}^2$, and $\lambda = \bar{x}^2$.

As an R^2 statistic, ρ is bounded between 0 and 1 ($\lambda \leq \kappa \leq \eta$) and is invariant to the scaling of test scores:

$$x'_a = \gamma x_a$$

$$\eta' = \frac{1}{N}\sum_{a \in A} \gamma^2 x_a^2 = \gamma^2 \eta$$

$$\kappa' = \frac{1}{N}\sum_{c \in C} N_c (\gamma \bar{x}_c)^2 = \gamma^2 \kappa$$

$$\lambda' = (\gamma \bar{x})^2 = \gamma^2 \lambda$$

$$\rho' = \frac{\gamma^2 \kappa - \gamma^2 \lambda}{\gamma^2 \eta - \gamma^2 \lambda} = \rho$$

This has two implications. First, if there is a change in the testing regime that preserves the general shape of the score distribution, then ρ is not mechanically affected. Second, cohorts that are more homogeneous (i.e., have prior test scores with a lower variance) do not necessarily have higher tracking measures, since the measure is conditional on the degree of variability in prior test scores.

Closely related to ρ is the measure used by Collins and Gan (2013) to study the impact of tracking on achievement in the Dallas Independent School District. The measure relates the

overall standard deviation of achievement within students' school-grade cohorts to the (enrollment-weighted) average standard deviation within students' classes:[40]

$$\alpha = \sqrt{\frac{\frac{1}{N}\sum_{a\in A}(x_a - \bar{x})^2}{\frac{1}{N}\sum_{c\in C}\sum_{a\in b_c}(x_a - \bar{x}_c)^2}}$$

A measure close to one suggests no sorting, while larger measures suggest more sorting by ability. When every class in a cohort has the same number of students, α is the following strictly positive monotonic transformation of ρ:[41]

$$\alpha = \sqrt{\frac{\eta - \lambda}{\eta - \kappa}} = \sqrt{\frac{1}{1-\rho}}$$

The relationship between these two is close to linear in the empirically relevant ranges of values, so that the choice to use one or the other is not consequential in our application.

B.2 Statistical Significance

In this section, we discuss different ways of determining whether a given estimate of our tracking measure is significantly different from zero. Since ρ is equivalent to the R^2 statistic from a regression of previous test scores on current class indicator variables, it is natural to consider an F-test of the joint significance of the class indicator variables. We calculate an F-statistic with degrees of freedom based on the number of students N and the number of class indicators N^C. Then, we generate a p-value from this F-statistic.

$$F = \frac{(\rho \,/\, N^C)}{\big((1 - \rho) \,/\, (N - N^C - 1)\big)}$$
$$p^F = 1 - F_{N^C, N-N^C-1}(F)$$

Since this test is based on large-sample asymptotic properties of the R^2 statistic, we interpret p^F as the probability a value as high as the observed ρ would be generated by repeated sampling from a large population of students. This thought experiment does not seem entirely appropriate to our setting, where we are trying to determine whether the degree to which a given set of students has been sorted is likely to have happened by chance.

For that reason, we also implement a finite sample method based on a different thought experiment: if a school randomly assigns a set of students A (with associated scores X) to a set of classes C, what is the probability that a value as high as the observed ρ would be generated? This is different from the repeated-sampling thought experiment above because the sets of students and classes (including class sizes) are fixed. Imagine repeatedly randomly assigning a cohort of students across their set of classes, and then for each permutation calculating the R^2 statistic, ρ^{ra}, from a regression of prior test scores on class indicator variables. Though we would ideally then calculate the fraction of simulated ρ^{ra} that fall above the actual value ρ, we implement an approximation that is more easily computed.

We derive a pseudo p-value based on the distribution of values ρ^{ra} takes under random assignment of students to classes. We first standardize ρ using the mean and standard deviation of ρ^{ra} across permutations:

[40] In our interpretation of the Collins and Gan (2013) measure, we weight the denominator by the number of students in each class, rather than weighting each class equally.

[41] We thank Edwin Leuven for initially pointing out this relationship to us.

$$\rho^Z = \frac{\rho - \rho^{ra,\mu}}{\rho^{ra,\sigma}}$$

Then, we calculate the p-value of that standardized measure using a t-distribution with degrees of freedom based on the numbers of students and classes:

$$p^Z = 1 - t_{N-N^C-1}(\rho^Z)$$

In this way, we can say how likely the observed level of tracking in the given school-grade-year would be if the school were not engaging in any kind of tracking.

Figure B1 compares p^F and p^Z, the p-values calculated from the F-test and from the random assignment counterfactual. They are highly correlated, but the former tends to give somewhat larger values. Figure B2 shows the distribution of ρ, with bins split into two based on whether the corresponding test would find ρ to be statistically significant at the 5% level. Both the F-test (top panel) and the random assignment counterfactual (bottom panel) find that larger values of ρ are more likely to be statistically significantly different from zero. Values of ρ beyond 0.15 are almost always statistically significant, regardless of the test.

It is worth noting that the mean of the distribution under random assignment, across permutations (indexed by $p \in P$), is a simple function of the number of classes N^C and the number of students N:

$$E_P(\eta) = \eta = \frac{1}{N}\sum_{a \in A} x_a^2 = E(x_a^2)$$

$$E_P(\lambda) = \lambda = \left(\frac{1}{N}\sum_{a \in A} x_a\right)^2 = \frac{1}{N^2}\sum_{a \in A} x_a^2 + \frac{1}{N^2}\sum_{a \in A}\sum_{j \neq a} x_a x_j = \frac{1}{N}E(x_a^2) + \frac{N-1}{N}E\left(x_a x_j | a \neq j\right)$$

$$E\left(x_a x_j | a \neq j\right) = \frac{N}{N-1}\lambda - \frac{1}{N-1}\,\eta$$

$$E_P(\kappa_p) = \frac{1}{N}\sum_{c \in C}\frac{1}{N_c}E_P\left(\left(\sum_{a \in b_c} x_a\right)^2\right) = \frac{1}{N}\sum_{c \in C}\frac{1}{N_c}E_P\left(\sum_{a \in b_c} x_a^2 + \sum_{a \in b_c}\sum_{j \neq a} x_a x_j\right)$$

$$= \frac{1}{N}\sum_{c \in C}\frac{1}{N_c}\left(N_c E(x_a^2) + N_c(N_c - 1)E\left(x_a x_j | a \neq j\right)\right)$$

$$= \frac{N^C}{N}E(x_a^2) + \frac{N - N^C}{N}E\left(x_a x_j | a \neq j\right) = \frac{N^C - 1}{N-1}\eta + \frac{N - N^C}{N-1}\lambda$$

$$\rho^{ra,\mu} = E_P\left(\frac{\kappa_p - \lambda}{\eta - \lambda}\right) = \frac{\left(\frac{N^C-1}{N-1}\eta + \frac{N-N^C}{N-1}\lambda\right) - \lambda}{\eta - \lambda} = \frac{N^C - 1}{N-1}$$

For that reason, rather than simulate $\rho^{ra,\mu}$ and $\rho^{ra,\sigma}$, we calculate these moments.[42]

B.3 Relative Tracking Measure

Our absolute measure of tracking ρ is affected by the distribution of class sizes. In this section, we develop an alternative measure that conditions on this. While reducing class size may be a tool to increase the degree of tracking and target instruction more closely to students'

[42] The formula for the standard deviation of the distribution of ρ^{ra} is more complex, but it is still a function only of the number and sizes of classes, the number of students, and moments of the distribution of previous test scores.

abilities, smaller classes may also be associated with increased resources or other policies unrelated to tracking. Our "relative" measure of tracking captures the portion of potential tracking (given the class size distribution) that is realized by the actual assignment of students to classes.

All else equal, if a grade has more classes, it will generally have a higher level of measured tracking ρ. Recalling that ρ is equivalent to an R^2 statistic from a regression of previous test scores on current class indicator variables, adding a class increases the number of explanatory variables by one. If a class with any previous test score variance is split in two, the R^2 will increase. The top panel of Figure B3 shows the distribution of $\rho^{ra,\mu}$, the mean of the unadjusted R^2 statistic under random assignment to classes, for cohorts with different levels of average class size. As expected, cohorts with the largest (and thus fewest) classes (quartile 4) have the smallest values.

Furthermore, measured tracking is affected by how the class size distribution interacts with the distribution of prior student achievement. Suppose that a cell of 120 students has 60 students with a score of 1 and 60 students with a score of 0. If two classes each have 30 students, and one has 60 students, then the students could theoretically be perfectly sorted into classes by previous test score. If all three classes have 40 students, there must be at least one class with both types of students. In this way, our unadjusted measure of tracking ρ is constrained by the set of classes into which students of differing achievement levels can be sorted.

To estimate the maximal achievable degree of sorting taking the class size distribution as given, we simulate the distribution of the R^2 statistic under strict assignment to classes according to prior achievement. In these strict assignment permutations, a class size is chosen at random from the set of available classes, and then the students with the highest previous test scores are assigned to fill the class. Next, another class size is chosen (without replacement), and the unassigned students with the highest previous test scores are assigned to that class. This continues until all classes have been chosen and all students have been assigned. Then, we calculate a counterfactual ρ^{strict} based on this assignment of students to classes. While we could take the mean across all possible permutations of class sizes, for simplicity we take the mean across 1,000 randomly selected permutations to calculate $\rho^{strict,\mu}$. The bottom panel of Figure B3 shows that there is a great deal of variation in the mean maximum achievable R^2, and that cohorts with the smallest (and thus most) classes (quartile 1) turn out to have the smallest values.

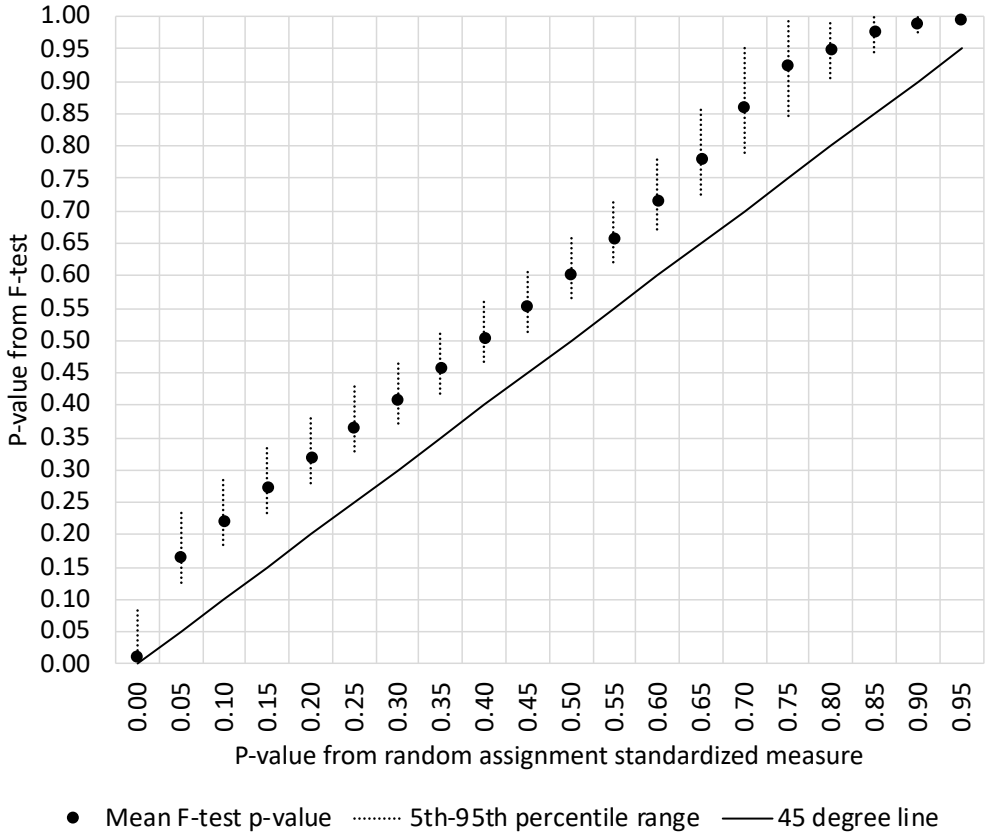
We construct our alternative relative measure of tracking as follows:

$$\rho^{rel} = \frac{\rho - \rho^{ra,\mu}}{\rho^{strict,\mu} - \rho^{ra,\mu}}$$

Interpreting the random assignment counterfactual as a lack of any tracking policy and the purposeful assignment counterfactual as the most intense tracking policy possible, this measure can be seen as the portion of possible tracking that is realized. The interpretation is loose: ρ^{rel} can be less than zero when the actual measure is below the mean simulated under random assignment, and it can be greater than one when the actual measure is above the mean simulated under purposeful assignment.

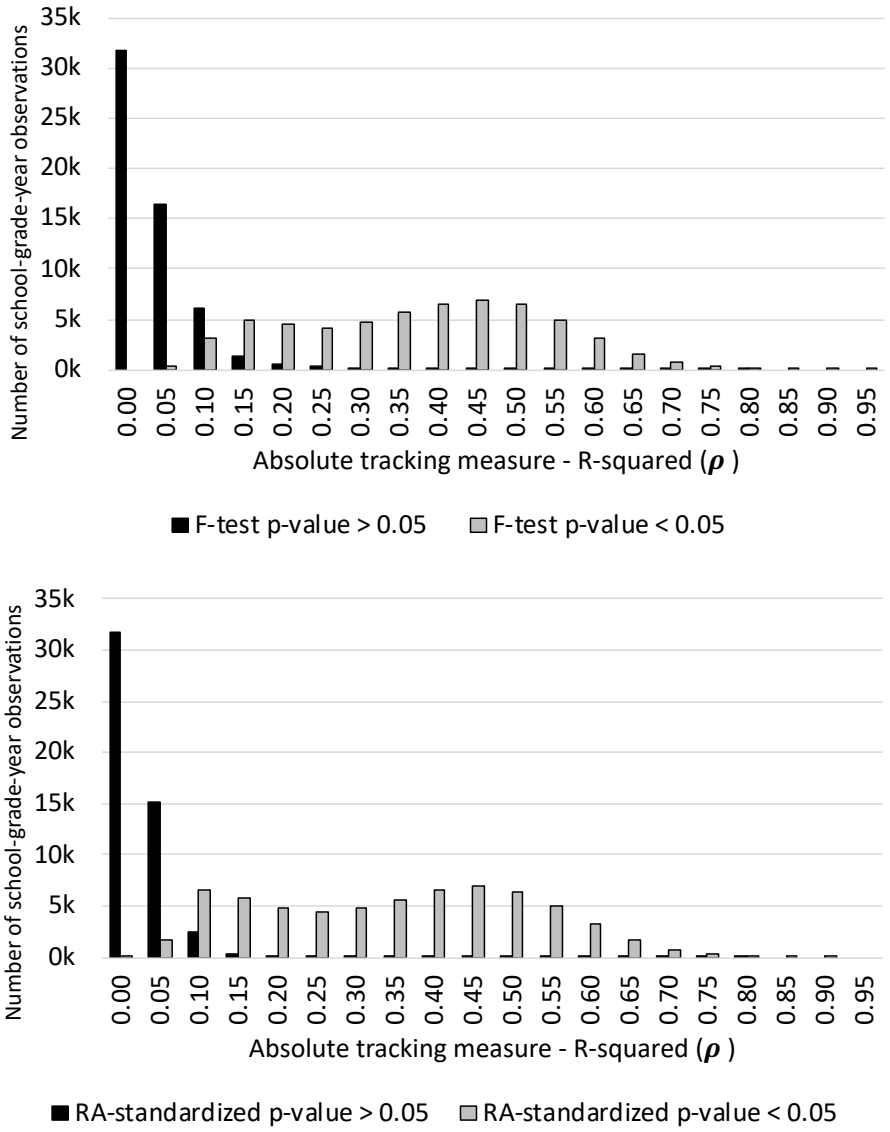
This measure is related to the "effective network isolation index" in Hellerstein et al. (2011). They standardize their index of network isolation (in the context of racial segregation) using the mean of that index from simulations with random assignment as well as the maximum value the index could take.

Figure B1. Comparison of P-values across Approaches



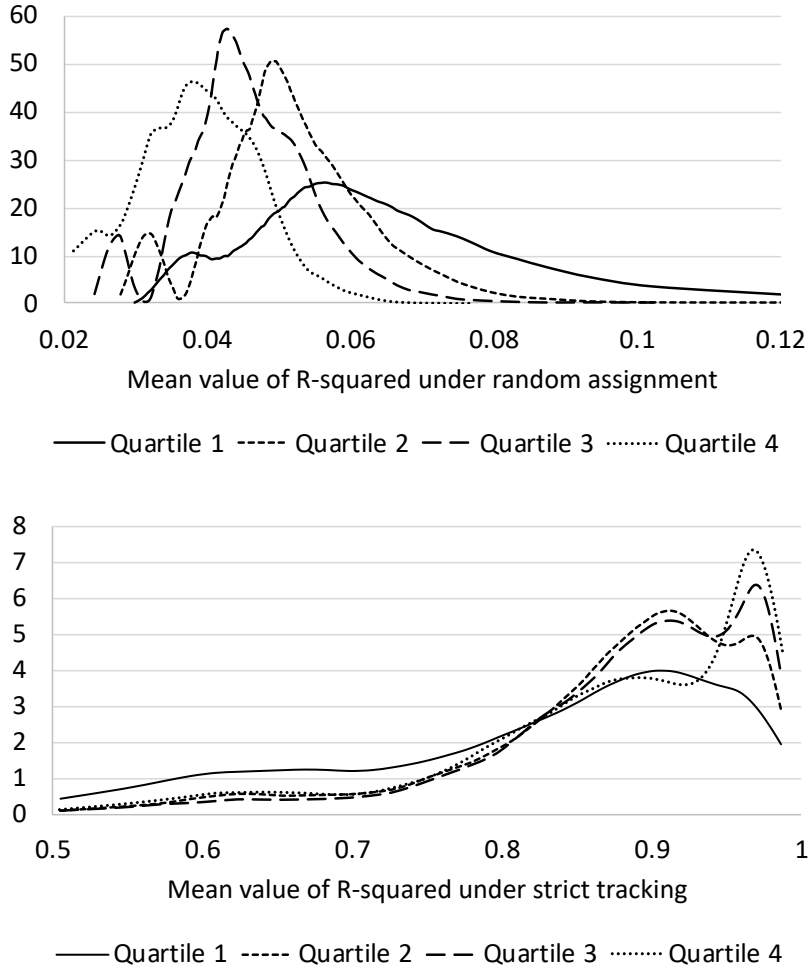
Notes: This figure compares the p-values from the F-test of the joint significance of the class indicators in the regression predicting prior achievement with those from the finite sample approach based on random assignment of students to classes. On the x-axis, the first bin is 0-0.05, the second bin is 0.05-0.10, and so on.

Figure B2. Level of Tracking by Confidence in Tracking, by Approach



Notes: This figure shows the number of school-grade-year observations for which the absolute tracking measure is (grey bars) and is not (black bars) statistically significant at the 5% level. In the top panel, statistical significance is based on a standard F-test. In the bottom panel, statistical significance is based on where the actual value falls in the distribution of values under random assignment of students to classes.

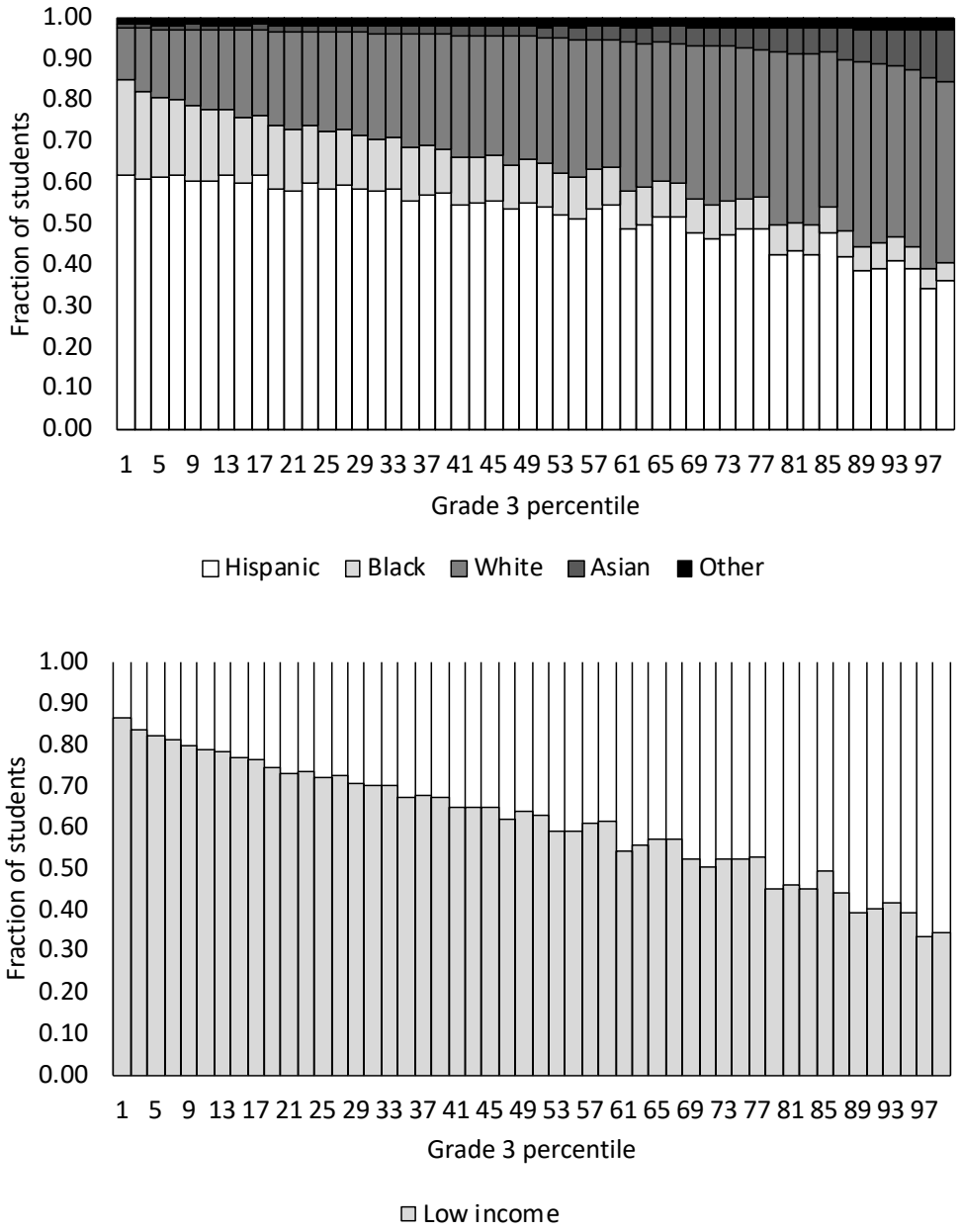
Figure B3. Distribution of R^2 under Random and Strict Assignment, by Average Class Size



Notes: The top panel shows the density of the mean R-squared value under random assignment to classrooms for the analysis sample of school-grade-years, while the bottom panel shows the density of the mean R-squared value under strict tracking by achievement. The quartiles are based on average math class size for the school-grade-year. Class sizes are on average 13, 17, 19 and 23 students moving from quartile 1 to quartile 4.

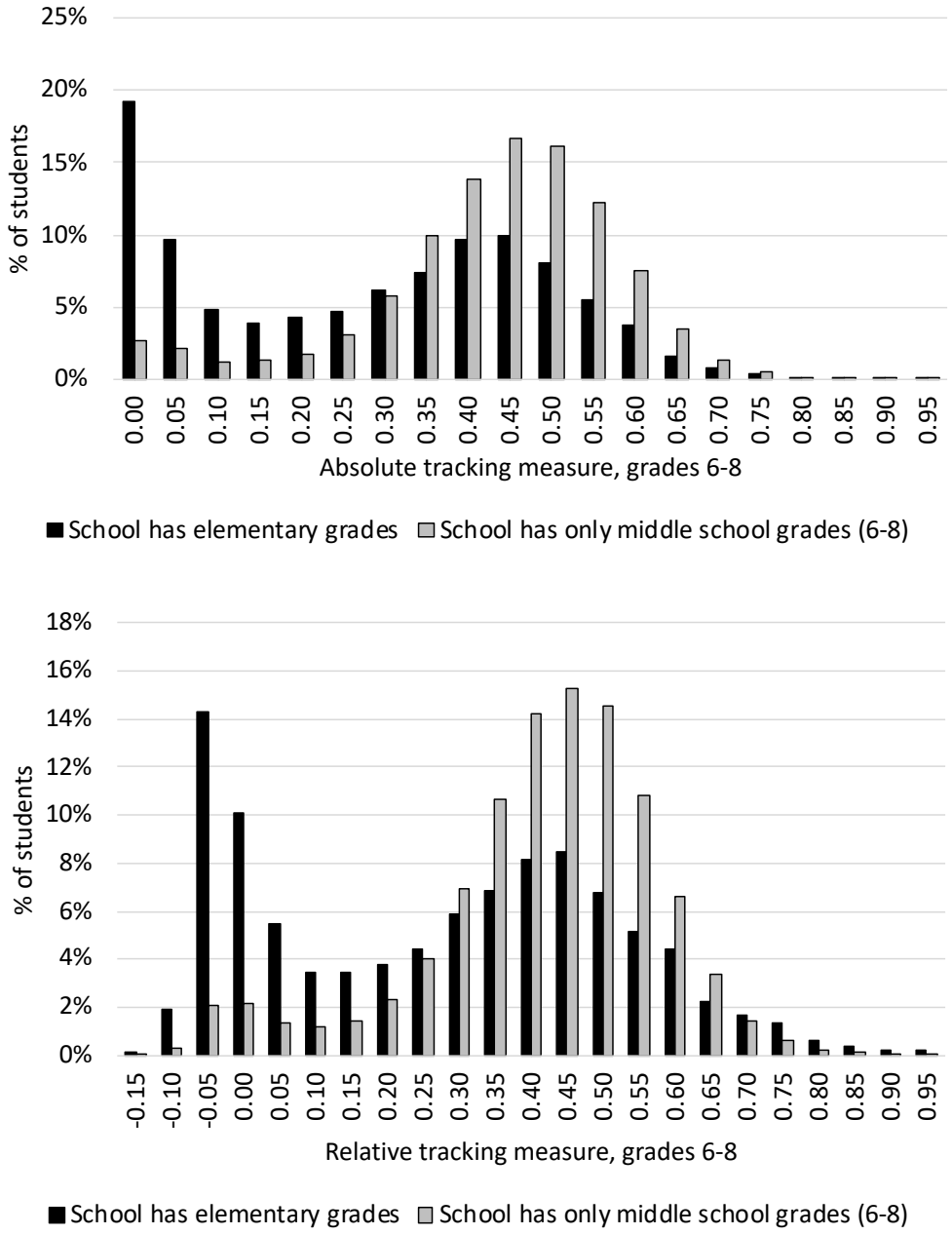
Appendix C. Supplementary Figures and Tables

Figure C1. Race/Ethnicity and Low-Income Shares, by Grade 3 Achievement Percentile



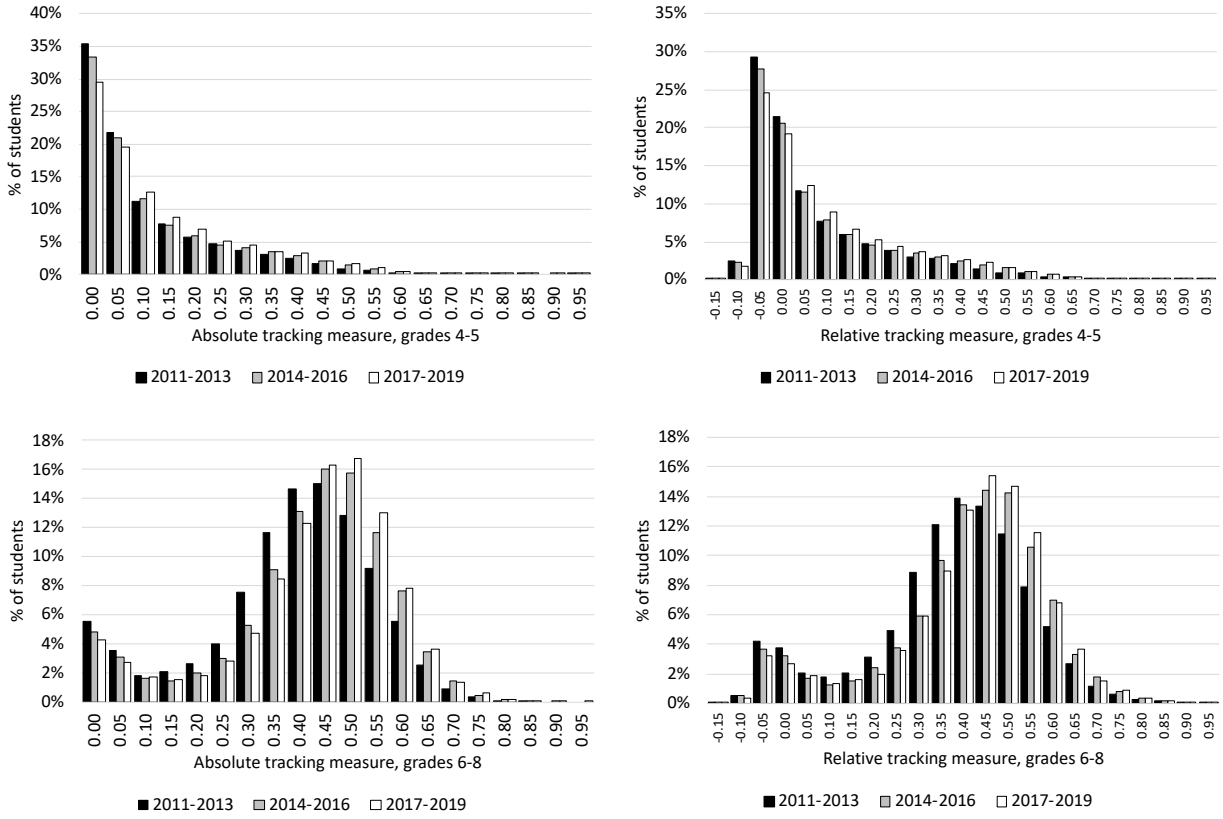
Notes: This figure shows race/ethnicity (top panel) and low-income (bottom panel) shares, by students' positions in the grade 3 math test score distribution. Low-income students are those who are eligible for free or reduced-price meals or certain public assistance programs (such as TANF).

Figure C2. Tracking Measures for Grades 6-8, by School Grade Composition



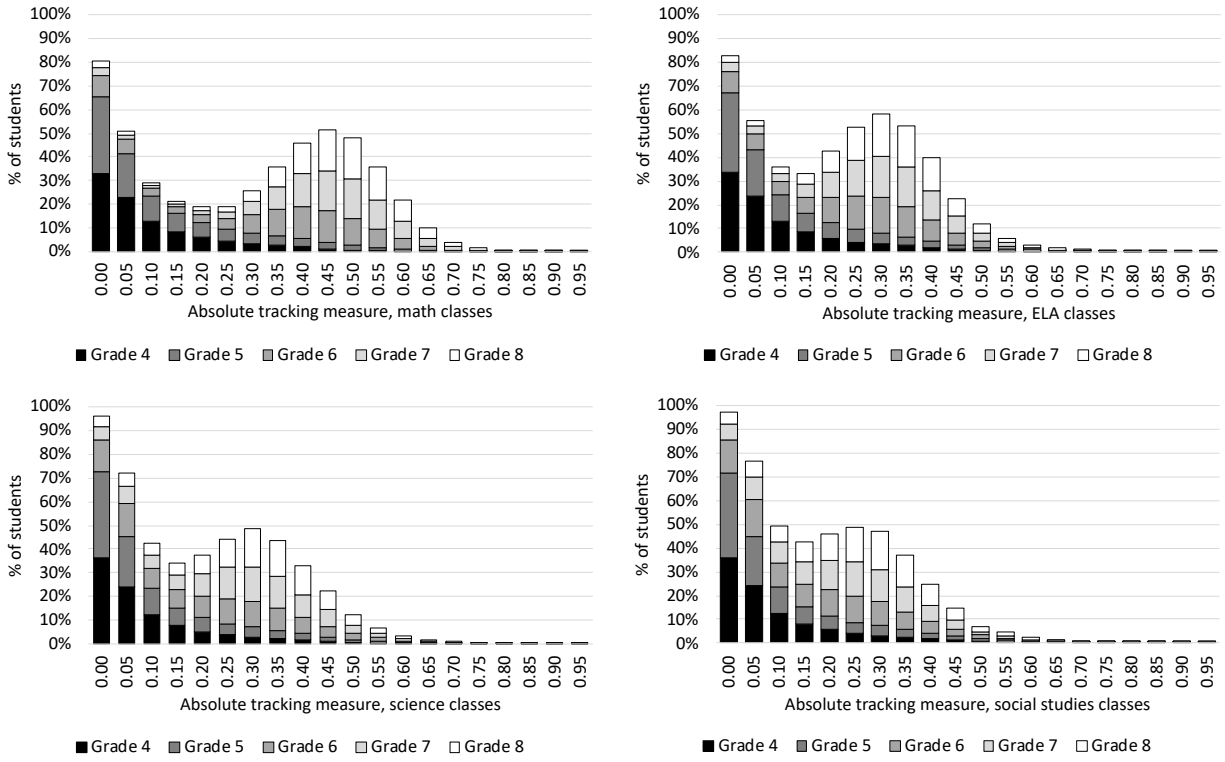
Notes: This figure shows the student-weighted distribution of the absolute (top panel) and relative (bottom panel) tracking measures for students in middle school grades (6-8), broken down by whether the school serves any grades below grade 6. Only a small share (14.6%) of middle school students is in schools with elementary grades.

Figure C3. Tracking over Time



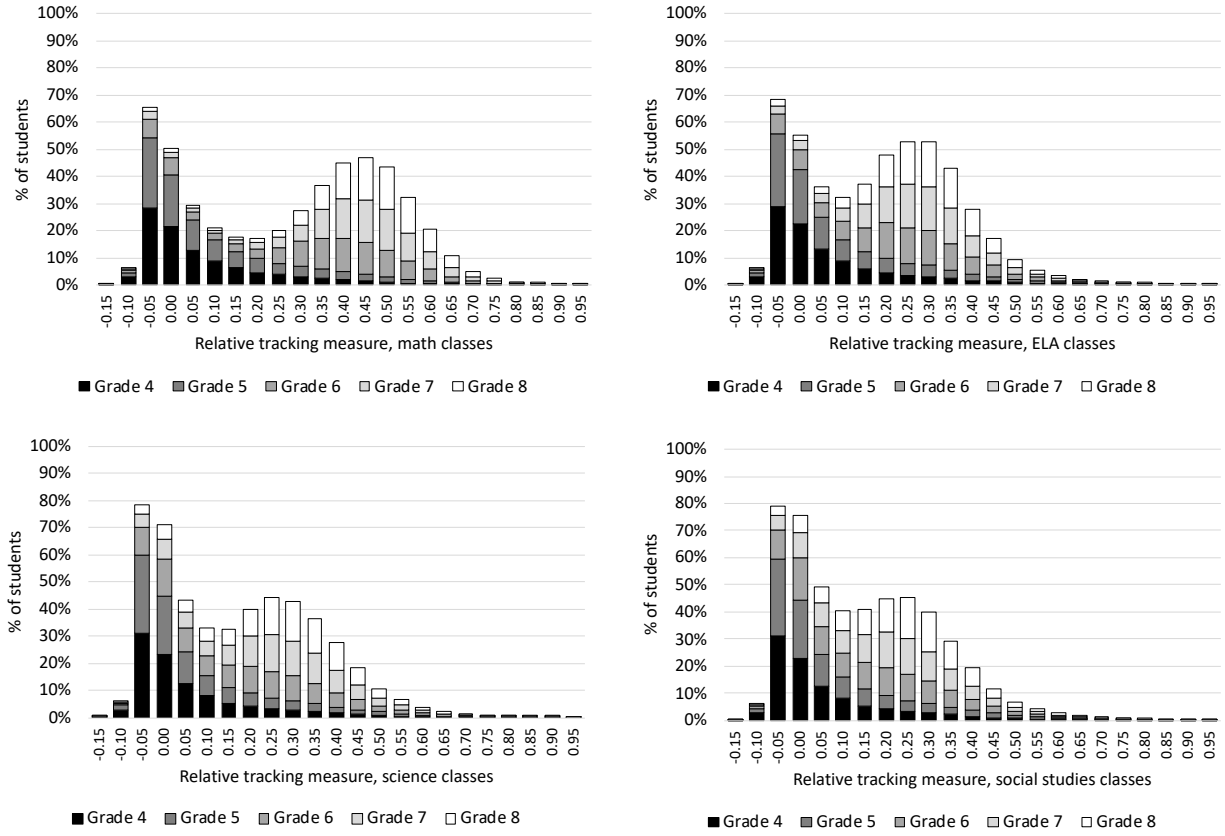
Notes: This figure shows the student-weighted distributions of the absolute and relative tracking measures, broken down by grade-level and time period.

Figure C4. Distribution of Absolute Tracking for Math and Other Subjects



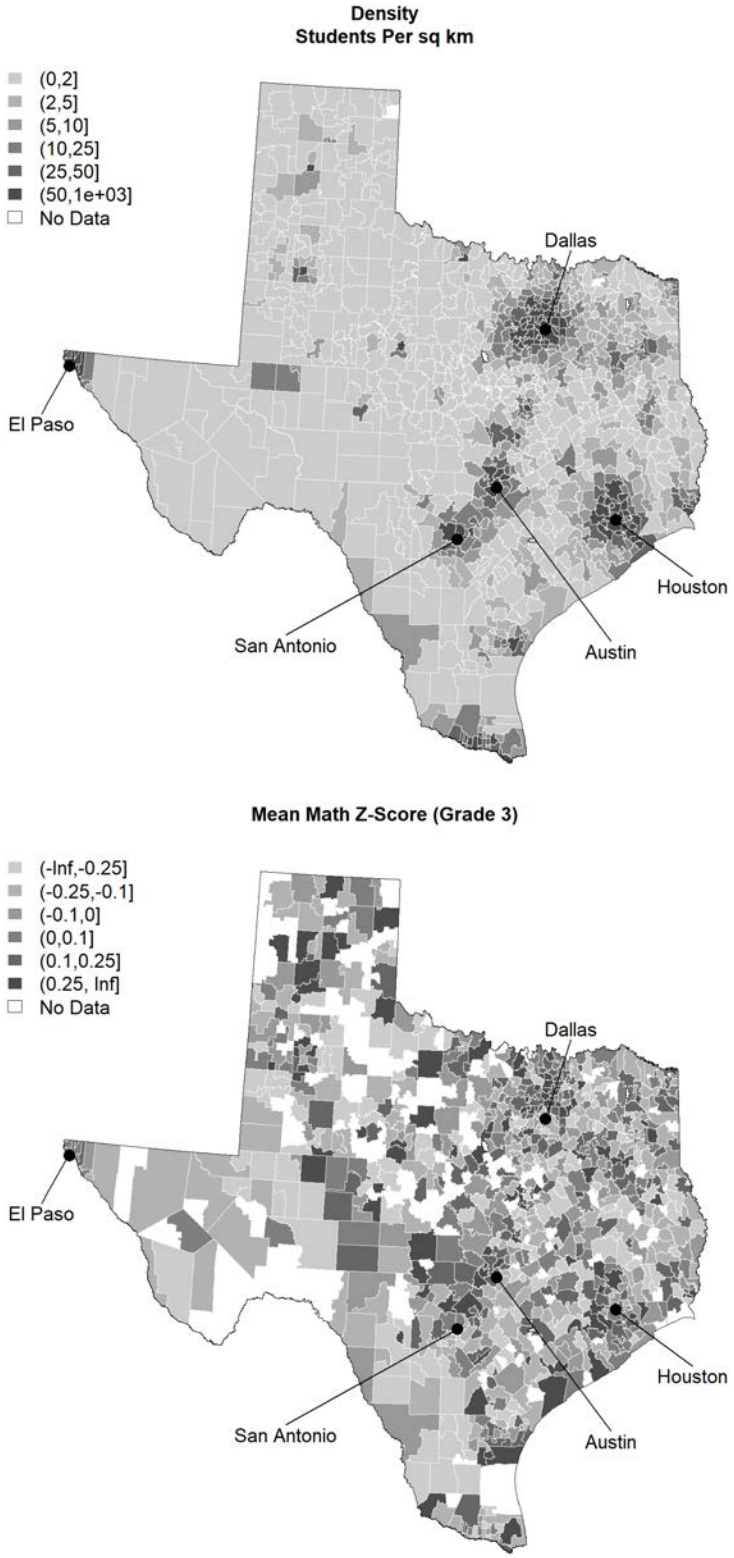
Notes: These panels show the student-weighted distributions of absolute tracking by prior math scores for math (top left), English language arts/reading (top right), science (bottom left), and social studies (bottom right) classes, broken down by grade.

Figure C5. Distribution of Relative Tracking for Math and Other Subjects



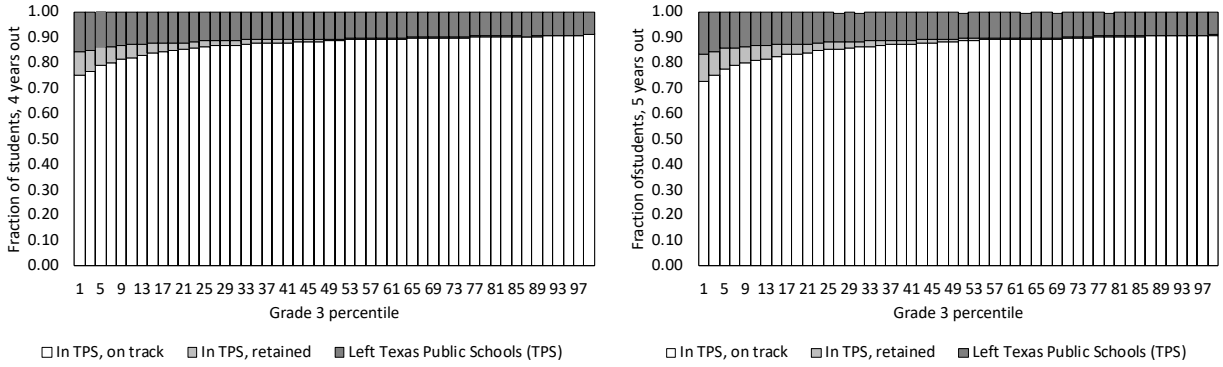
Notes: These panels show the student-weighted distributions of relative tracking by prior math scores for math (top left), English language arts/reading (top right), science (bottom left), and social studies (bottom right) classes, broken down by grade.

Figure C6. Population Density and Achievement Levels Across Districts



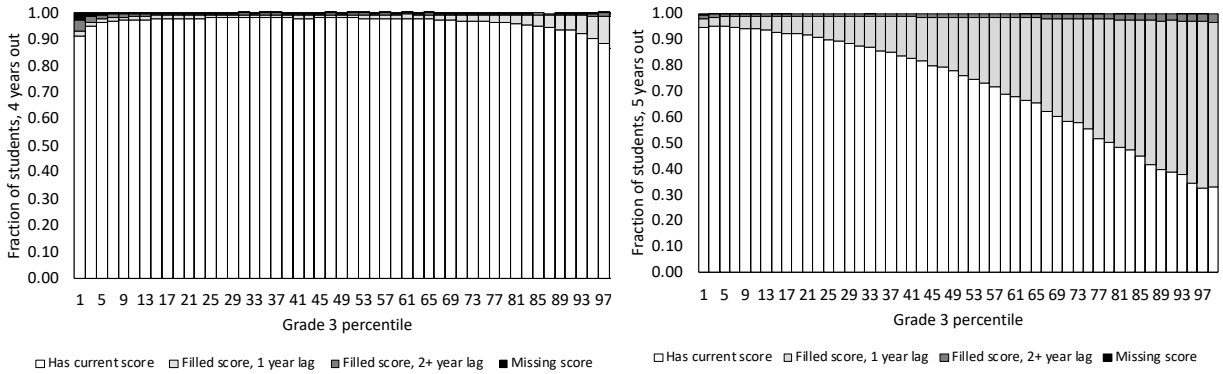
Notes: The maps show geographic variation in population density (top panel) and 3rd grade achievement levels (bottom panel) across school districts in Texas.

Figure C7. Enrollment Status 4 and 5 Years Out, by Grade 3 Achievement Percentile



Notes: The bars show the fraction of students that has left the Texas Public Schools (darkest bars) and the fractions enrolled in the expected grade (lightest bars) or in a grade below that expected (intermediate bars), by students' positions in the grade 3 math test score distribution. The left (right) panel shows these statistics for 4 (5) years after grade 3.

Figure C8. Test Score Patterns 4 and 5 Years Out, by Grade 3 Achievement Percentile



Notes: From lighted to darkest, the bars show the fraction of enrolled students that has current math scores and the fractions with no current score but with a percentile score filled in from the prior year, a percentile score filled in from two or more years ago, and no available score since grade 3. The left (right) panel shows these statistics for 4 (5) years after grade 3.

References for Appendices

Collins, C. A., & Gan, L. (2013). Does sorting students improve scores? An analysis of class composition. National Bureau of Economic Research Working Paper No. 18848.

Hellerstein, J. K., McInerney, M., & Neumark, D. (2011). Neighbors and coworkers: The importance of residential labor market networks. *Journal of Labor Economics*, 29(4), 659–695.

International Association for the Evaluation of Educational Achievement (IEA) (2015). "Trends in International Mathematics and Science Study (TIMSS)." Retrieved from NCES International Data Explorer (https://nces.ed.gov/surveys/international/ide/) (August 2, 2020).

National Center for Education Statistics (NCES) (1990, 1992, 1996, 2000, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, and 2019). National Assessment of Educational Progress (NAEP) Mathematics Assessments. U.S. Department of Education, Institute of Education Sciences. Retrieved from https://www.nationsreportcard.gov/ndecore/xplore/nde (August 3, 2020).