CESifo WORKING PAPERS

9928
2022

September 2022

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

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

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9928*

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

Abstract

Schools often track students to classes based on ability. Proponents of tracking argue it is a low-cost tool to improve learning since instruction is more effective when students are more homogeneous, while opponents argue it exacerbates initial differences in opportunities without strong evidence of efficacy. In fact, little is known about the pervasiveness or determinants of ability tracking in the US. To fill this gap, we use detailed administrative data from Texas to estimate the extent of tracking within schools for grades 4 through 8 over the years 2011-2019. We find substantial tracking; tracking within schools overwhelms any sorting by ability that takes place across schools. The most important determinant of tracking is heterogeneity in student ability, and schools operationalize tracking through the classification of students into categories such as gifted and disabled and curricular differentiation. When we examine how tracking changes in response to educational policies, we see that schools decrease tracking in response to accountability pressures. Finally, when we explore how exposure to tracking correlates with student mobility in the achievement distribution, we find positive effects on high-achieving students with no negative effects on low-achieving students, suggesting that tracking may increase inequality by raising the ceiling.

JEL-Codes: H750, I210, I240, I280.

Keywords: ability tracking, achievement mobility.

Kate Antonovics
University of California
San Diego / CA / USA
kantonov@ucsd.edu

Sandra E. Black
Columbia University
New York / NY / USA
sb4338@columbia.edu

Julie Berry Cullen
University of California
San Diego / CA / USA
jbcullen@ucsd.edu

Akiva Yonah Meiselman
Department of Economics
University of Texas at Austin / USA
yonah.meiselman@gmail.com

August 2022
The authors are grateful to Mindie Hsu, Eli Mogel, Anjali Priya, Kelly Wang, and Sonia Yan for excellent research assistance. The conclusions of this research do not necessarily reflect the opinion or official position of the Texas Education Research Center, the Texas Education Agency, or the State of Texas. 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.

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

[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).

suggest that tracking by ability across classes is prevalent. 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 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 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 this. 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

[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.
[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 class-level tracking measures, our measures miss the extent to which students are sorted into different ability groups within the same classroom.

tracking in the grand scheme of student sorting? What are the explicit and implicit mechanisms by which schools track students? Which districts and schools track students to a greater degree? Finally, we consider the impact of exposure to more tracked regimes on future achievement for students 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, because school assignment is based primarily on residential location, much of the sorting takes place across districts and schools. In fact, only 10% and 17% of the variation in prior scores (within grade-years) is explained by districts and schools, while 44% is explained by classes.[9] 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 schools and grades. Consistent with national survey data, we find that middle 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 10% (37%) of its potential to track students across classes, this ranges from no tracking at the 5th percentile (both elementary and middle schools) to 42% (73%) at the 95th percentile.

Schools can facilitate tracking in a number of ways, including by establishing differentiated curricula for advanced and remedial students and more aggressively classifying students for gifted and special education programs. When we examine the decisions that are most predictive of tracking, we find that schools appear to operationalize tracking through the classification of students into categories such as gifted and disabled, as well as through curricular differentiation. This is true even after controlling extensively for the distribution of student

[9] See Appendix Table C1 for details.

ability and, to a more limited extent, student socioeconomic status.

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. And when we examine the relationship between accountability pressure and tracking, our results suggest that schools *reduce* tracking concurrent with receiving a low performance rating.

Finally, 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 is not related to future test score growth. For those initially at the top, however, exposure to more tracking is beneficial. For example, our results suggest that a one standard deviation increase in exposure to middle-school tracking would lead to a 1.3 percentile increase in predicted test scores 5 years after 3rd grade for students initially at the 75th percentile. These findings are consistent with tracking aggravating inequities in educational outcomes, but primarily by benefitting those already at the top.

To examine possible mechanisms, we use a similar empirical strategy to examine how tracking relates to the average class size and peer quality experienced by students at different points in the initial test score distribution. Not surprisingly, we find that students who are

exposed to more tracking face more inequality in average peer achievement: low-achieving students experience poorer-achieving peers while high-achieving students experience higher-achieving peers. 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.[10]

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). 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.[11]

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. With respect to policy determinants, other work has studied how policies such as "algebra for all" affect tracking (e.g., Domina et al. 2016), and hypothesized how school accountability would affect tracking (e.g., Fu and Mehta 2018). As far as we know, we are the first to study the latter question empirically.

[10] 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).

[11] 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).

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.[12] For example, Card and Giuliano (2016) and Cohodes (2020) use regression discontinuity designs and 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. To examine the relationship between tracking and test score mobility in Texas, we rely on across-cohort variation for identification and apply methods consistent with recent work by Reardon (2019), who uses administrative test score data to document patterns of achievement gains across grades for US school districts.

The paper unfolds as follows. In the next section, we discuss the data and methodology. 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 (ERC). These data cover the universe of public

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

elementary and secondary school students in Texas and enable us to link students to classes and courses 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 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.

The fact 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 are key barriers to measuring tracking past grade 8. 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. 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 are able to identify a focal math class. We include all school-grade-years from 2011 to 2019 with at least two classes with two or more students from the tracking sample.

Table 1 presents summary statistics by grade for our sample of school-grade-years; all statistics are student-weighted. Our sample represents over 4,000 elementary and 2,000 middle schools across 1,000 districts. It represents 96% of students enrolled in grades 4-8 in Texas over our period.[15]

---

[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] One percent of student-year observations are in generic courses (mostly in grades 4-5), and 0.3% are taking multiple math courses (mostly in grades 7-8).

[15] The shares of students without a focal math course and missing prior test scores range from 4-10% and 6-7% across grades, respectively. Exam scores may be missing for idiosyncratic reasons, such as student absence or migration, but also for systematic reasons due to policies. 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

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 level of the school-grade-year cell.

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,

STAAR regime, students enrolled in grades 3-8 take an end-of-course (EOC) assessment rather than the grade-level math assessment if they are receiving instruction in a high school level course for which an EOC assessment exists (e.g., algebra).

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

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 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 ($\rho^{strict,\mu}$):[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, 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

[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.

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 size, which may be driven by other considerations and have its own impact on outcomes.

3. Scope and Nature of Tracking

In this section, we first present our findings on the degree of sorting by ability across classes within a school and how that compares to sorting at other levels, such as across schools within a district, and sorting on other dimensions. We then explore the potential mechanisms schools may use to track students, such as through curriculum differentiation and special instructional programs.

3.1 Scope of Tracking

Figure 1 shows the student-weighted distributions of the absolute and relative tracking measures. Across all grades, the mean level of absolute tracking is 0.23 (with a standard of deviation 0.28), implying that the average student is in a cohort where class assignments explain 23% 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 0.21, and the standard deviation is 0.25. Using our loose interpretation of relative tracking, this suggests that on average 21% 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 provide a sense of how important within-school tracking is relative to sorting across schools, Figure 2 shows the distribution of absolute tracking across classes within a school in gray, while the black bars show the distribution of absolute tracking when calculated across schools within a district, capturing across-school ability sorting. As is clear, across-school sorting – after residential and school choices are made and before students arrive in the classroom – is much lower than sorting within schools. There is a large spike near zero, and it is rare for across-school sorting to explain more than 20% of the variation in prior scores.

Figure 3 also makes it apparent that, across classes within schools, there is much less sorting by race/ethnicity and socioeconomic status than by prior test scores.[20] The figure shows the distributions of our absolute and relative tracking measures when race/ethnicity (i.e., Black or Hispanic vs. non-Black and non-Hispanic) or SES (i.e., eligible vs. ineligible for subsidized meals) is used in place of prior achievement. Compared to our measures of ability tracking, these race and SES tracking measures are much more tightly clustered around the no-tracking benchmarks, underscoring that our ability tracking measures are not simply proxies for other types of across-class sorting.

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 middle school grades.[21] This pattern helps to explain the bimodality observed for within-school tracking in the preceding figures. It is also expected since sorting by ability will rise as

[20] Perhaps not surprisingly, there is relatively more sorting along these demographic dimensions across schools within a district compared to across classes within a school, particularly by race/ethnicity (See Appendix Table C1).
[21] Appendix Figure C1 shows that the grade configuration also matters, in that middle school cohorts served in schools that also have elementary grades are less tracked. Figure C2 shows that tracking increases slightly across years in our sample period.

students begin to take courses that are differentiated not only by level of difficulty and pace but also by subject content. In addition to the differences across grade levels, the figure reveals substantial variation in tracking within grade levels. The fraction of potential tracking realized ranges from none at the 5th percentile to 42% at the 95th percentile for elementary school students, and from none at the 5th percentile to 73% at the 95th percentile for middle school students.

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 recalculate our tracking measures for English language arts/reading, science, and social studies classes. We continue to use prior math scores as the proxy for student achievement, so these measures capture how sorted students are according to math ability in non-math classes and are readily comparable to our baseline measures. Table 2 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.[22] Thus, any given degree of sorting across math classes translates to a greater degree of sorting throughout the school day in the elementary grades.

3.2 Nature of Tracking

To provide a sense of how coordinated and purposeful tracking policy is, we first 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 results in Table 3 reveal that 69% of the variation in our absolute tracking measure is explained by district-grade-year fixed effects,

[22] These correlations are likely understated due to measurement error. Appendix Figure C3 shows the distributions of the absolute tracking measure for all four core subjects for visual comparisons.

suggesting that the district plays a substantial role in setting policies that affect tracking. When we focus exclusively on larger districts with at least 6 schools for every grade across all the years in our sample, the fraction of the variation accounted for by district-grade-year fixed effects falls to 36%. Nonetheless, it is clear that districts matter for tracking policy. When we break down our results by grade, we find that the district plays an especially important role in middle school, where district-grade-year fixed effects account for 79% of the variation in tracking.[23]

Next, we examine the different ways in which schools sort students across math classes within a school. The assignment of students to classrooms based on ability could 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 and talented students, English learners, and students with disabilities, 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 a variety school-grade-year characteristics that are intended to capture the programming choices that could be correlated with tracking, taking as

[23] Further suggestive evidence that tracking practices are intentional is the persistence of tracking across time. Carrying out the same type of exercise by including school-grade indicators without the time component, we find that 60% (54%) of the variation in absolute (relative) tracking is explained, with the remaining variation over time across school-grades.

given the student achievement distribution. Table 4 presents the results.[24] Across all specifications, we include controls for student prior achievement, grade level and configuration, cohort and district size, district property wealth, type of locale, and year. Observations are weighted by school-grade-year enrollment, and standard errors are clustered at the district level.

To identify the role of school policies that may lead to segregation of low- and high-achieving students, we include controls for the shares of students in a variety of special needs categories and programs. For non-English-speaking students, we include the shares of students receiving instruction in core subjects other than English in more isolated settings (i.e., bilingual non-two-way and ESL content-based classes) and the shares in settings where they are integrated with other students (i.e., bilingual two-way and ESL pull-out classes). We also control for the fraction of students classified with physical and other disabilities, as well as the fraction of classified as gifted.

Moving across the columns, additional variables are successively added to the control set. These include controls for resource levels and math curricular differentiation (column 2), for the tails of the student prior achievement distribution (column 3), district fixed effects (column 4), and school fixed effects (column 5). Our measure of curricular differentiation captures the dispersion of students across different math courses and is equal to one minus the Herfindahl index of 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.

We focus on the results in column 5, which isolate within-school variation and control flexibly for the distribution of prior student achievement. These results reveal that tracking appears to be operationalized through more aggressive classification of students. We see that

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

cohorts that are more tracked have higher shares classified both as gifted and disabled. The link to disability shares appears only for the nonphysical disability categories, which are dominated by emotional and learning disabilities and more subject to discretion in classification. Any link to physical disabilities would more be more likely to reflect the student case-mix. Other signs of willingness to segregate students according to needs – such as serving English learners in more isolated settings – are not significantly related to greater tracking. Turning to other programmatic variables, we find that cohorts that are more tracked have access to greater resources, including more experienced teachers and smaller classes. And, not surprisingly, greater math curricular differentiation is associated with greater tracking.

Overall, the results in this section highlight that our measures of tracking reflect bundles of district and school instructional policies and practices.

4. Determinants of Tracking

In the previous section, we examined how schools operationalize tracking. In this section, we first examine how a variety of local characteristics predict the degree of tracking within a school-grade-year cohort. While this exercise is descriptive in nature, we follow by estimating plausibly causal impacts of external pressure via the statewide accountability system.

4.1 Local Determinants

Different schools and districts are likely to perceive the possible equity vs. efficiency tradeoffs involved with tracking differently, depending on their constituencies. For example, schools serving students with wide disparities in ability might see more instructional benefits to sorting by ability across classes. 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.

To quantify this, Table 5 presents the results from regressing our school-grade-year absolute tracking measure on various school, district, and county characteristics.[25] As in the previous analysis, all specifications include controls for student prior achievement, grade level and configuration, cohort and district size, district 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 6 show the sensitivity of our results as we add different sets of covariates, with column 5 including district fixed effects and column 6 including school fixed effects.

Consistent with previous studies, column 1 indicates that tracking is positively and statistically significantly associated with mean lagged test scores, implying that schools that serve high-achieving students tend to track more. In column 2, however, we see that once we control for the variability of test scores, as measured by the standard deviation of lagged test scores within a school-grade-year, the coefficient on mean lagged tests scores becomes negative and statistically significant. We also find that the standard deviation of lagged math test scores is a positive and statistically significant predictor of tracking. Recall that the standard deviation of test scores is *not* mechanically related to our measure of tracking, suggesting that the perceived net benefits of tracking are increasing with the heterogeneity of student ability. The relationship between tracking and the mean and standard deviation of a school's lagged test scores becomes less pronounced in columns 3 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

[25] In results not shown, we find qualitatively similar results for the relative measure.

demographics—as proxied by the racial composition and shares of students who are low income and limited English proficient—and the degree of tracking.

Turning to variables related to schooling options, the results in Table 5 suggest tracking is higher at magnet schools and lower at charter schools. 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 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 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.[26]

4.2 Tracking and School Accountability Pressure

Since tracking policies are highly decentralized, we know very little about how schools and districts might adjust tracking in response to state and federal education policies. Policies such as funding levels, minimum class size requirements, etc., likely induce shifts in how schools

[26] 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).

organize instruction. Here, we consider the effects of school accountability schemes that require a certain proportion of students to achieve satisfactory performance; these schemes create incentives for schools to focus on the more marginal students to improve passing rates. As a result of being identified as underperforming, schools might group these students to better target them, resulting in an increase in tracking, or might alternatively reduce tracking to increase exposure to higher-performing peers.

There is evidence that accountability pressures lead to differential gains across the distribution of prior achievement in response to accountability pressure.[27] For example, in the Texas context, Reback (2008) finds that achievement gains are largest for students whose gains have the greatest marginal impacts on their schools' ratings. Similarly, in Chicago, Neal and Schanzenbach (2010) find learning gains are concentrated among "bubble" students in the middle of the distribution who have a reasonable chance of becoming proficient. Just what instructional or allocational changes lead to these differential gains has been less well-identified.

Though we are not aware of any evidence on tracking, there are a few studies finding ties between accountability and within-class ability grouping. Using data from a nationwide survey of teachers along with focus groups and in-depth interviews with teachers, Bradbury (2018) shows that the introduction of a statutory assessment in England was associated with an increase in the pressure felt by teachers to group students by ability, despite their uncertainty about the appropriateness of this grouping. Using nationally representative data from the US, Reback et al. (2014) finds that accountability pressure from the federal No Child Left Behind Act leads teachers to shift time away from whole-class instruction.

To explore the relationship between accountability pressures and tracking, we consider

[27] See Figlio and Deming (2016) for an overview of the broader impacts of school accountability.

the accountability system that was in place in Texas for the years 2013-2017 and focus on the receipt of low "unacceptable" ratings. At the start of this period, the accountability system newly emphasized learning gains, thus exposing schools with high achievement levels but low progress to the risk of being sanctioned.

In our analysis, we exclude schools that were recently "treated," in that they received a low rating in the years leading up to our sample period, during 2008-2011.[28] For the remaining schools, we identify whether the school received a low performance rating between 2013 and 2017. We then estimate the relationship between the timing of this low performance rating and the degree of tracking at the school using an event-study framework. Specifically, we regress our school-grade-year absolute tracking measure on indicators for the year that the school was given a low performance rating along with lags and leads (and controls for student and school characteristics). The coefficients thus describe the time pattern of tracking relative to the timing of the sanction, with two years prior as the omitted category.

Table 6 presents these results. Though we would expect to see changes starting in the following year if the change were in response to the rating per se, we see changes in tracking in the year of the low performance rating. This may suggest that schools respond to the risk of low ratings and begin to reorganize instruction prior to actually receiving the low rating. Importantly, we find that tracking actually decreases with these low performance ratings, which is quite different from the findings in the previous research on ability grouping.[29]

5. Implications of Tracking

Given the prevalence of tracking, a fundamental question is how it affects student

[28] No schools received ratings in 2012 during the transition.
[29] To address potential biases due to heterogeneous treatment effects, we also followed the method developed in Callaway and Sant'Anna (2021). While the estimates are somewhat attenuated, the conclusions are unchanged.

academic performance, and how this varies across the achievement distribution. We also want to understand how tracking impacts the educational environment for students at different points in the achievement distribution. 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 a tracking measure) between 2011 and 2015. We characterize students by their percentiles in the year-specific statewide grade 3 math test score distribution. We then evaluate the distribution of outcomes for up to 5 years after grade 3 (which, for most students, is grade 8).[30]

In each year that students are enrolled in the Texas Public Schools, we observe the level of tracking they are exposed to in that school-grade-year. We also observe their math test scores (which we convert to year-specific statewide percentiles) and their math classes.[31] For their math classes, we observe the grade-level of the subject, as well as class size and peer quality. Peer quality is proxied by the average math test z-scores of classmates, calculated based on either prior year test scores or initial test scores from the first time each student is observed with a non-missing score.[32]

5.1 Test Score Mobility

To examine test score 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

[30] To understand how this affects sample selection, Appendix Figure C4 shows the share enrolled 4 and 5 years out across the distribution of grade 3 test scores.
[31] Appendix Figure C5 shows that 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. 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.
[32] Class size and peer quality are both calculated using all students enrolled in the class, regardless of whether they are in the longitudinal sample or not.

score distribution several years later.[33] This allows us to assess the relationship between tracking and test score mobility for students at different points in the distribution of initial test scores.

Figure 5 shows the relationship between a student's initial position and their percentile rank 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 absolute tracking, based on average exposure across grades.[34] Students exposed to more tracking experience higher test score growth at almost all points of the distribution at both time frames. Of course, these cross-sectional patterns do not necessarily reflect a causal relationship since test score growth could be impacted by a variety of factors 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 generate parametric estimates of mobility at the 25th and 75th percentiles of the initial test score distribution 2 through 5 years after grade 3. We use a parametric approach, as the data can be sparse at the cohort level. For each (school-year-grade 4) cohort, we regress the *t* years later percentile on the grade 3 percentile separately for students above and below the statewide median of the grade 3 test score distribution. We use the estimated coefficients to predict the position *t* years after grade 3 at the 25th and 75th percentiles *p* for cohort *c* from school *s*: $\hat{Y}_{pcs}^{3+t}$.

With these measures in hand, we estimate the following school-cohort level regressions for the various time horizons, weighted by the number of students in each school-cohort:

$$\hat{Y}_{pcs}^{3+t} = \beta_0 + \beta_2 T_{cs}^{4-5} + \beta_3 T_{cs}^{6-8} + \beta_6 X_{cs}^{3} + \alpha_s + \delta_c + \epsilon_{pcs},$$

[33]This is also similar to the income mobility literature (e.g., Chetty et al. 2014; Hashim et al. 2020), which relates parents' position in the income or education distribution to their child's position.
[34] We first calculate the student-weighted average of tracking over a school-cohort in each year since grade 3, and then take the simple average across years.

where T_{cs}^{4-5} is the average school-cohort tracking exposure in grades 4 and 5 and T_{cs}^{6-8} is the average school-cohort tracking exposure in grades 6-8.[35] Separating 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. X_{cs}^{3} includes the mean and standard deviation of grade 3 standardized test scores for the school-cohort to control for the initial distribution of ability. All regressions also include school and cohort fixed effects, as well as controls for the fraction of students in the school-cohort who have enrollment records 2 through 5 years after grade 3. Thus, the coefficients on tracking exposure are identified from variation across cohorts within schools over time, holding constant initial school-cohort ability and enrollment patterns.

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 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 who advance one grade each year. This additionally helps address measurement error. Lingering concerns with interpreting our estimates as the causal impact of tracking on test score growth are that changes in district-level tracking may coincide with other policy changes that impact student test scores.

Table 7 presents our ordinary least-squares (OLS) and instrumental variables (IV) results. Each pair of elementary and middle school tracking exposure coefficient estimates is from a

[35] We take the mean of tracking 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-grade means across the relevant years since grade 3 as defined by students who progress normally.

separate regression. Moving down the rows, the number of years since grade 3 increases from 2 to 5. The first four columns are based on our absolute measure of tracking, while the second four are for relative tracking.

We first look at the effects of tracking in elementary and middle school on the predicted performance of students only 2 years after grade 3, when most students are in grade 5. Reassuringly, we do not see any impacts of upcoming exposure to middle-school tracking. Exposure to more tracking in elementary school reduces predicted performance two years later for lower-achieving students but has no statistically significant impact on higher-achieving students. The negative impacts are larger in the IV specifications that rely on variation in tracking across district-cohorts.

For later outcomes (3, 4 and 5 years after grade 3), exposure to tracking at either grade level has little effect on the later performance of lower-achieving 3rd graders. The statistically insignificant point estimates are slightly negative for elementary school tracking exposure and slightly positive for middle school tracking exposure. Among higher-achieving 3rd graders, however, we see clear benefits associated with exposure to tracking in middle school, with no effects of exposure in elementary school. The magnitude of the estimates suggests that a one standard deviation increase in exposure to middle-school tracking (which is 0.10 relative to a mean of 0.43 for our absolute measure) would lead to a 1.1 percentile increase in predicted test scores 4 years after grade 3 and a 1.3 percentile increase in predicted test scores 5 years after grade 3 for students initially at the 75th percentile.

5.2 Distribution of Educational Inputs

We next investigate how peer quality, class size, and curricular progress vary for students at different points in the initial achievement distribution in more versus less tracked regimes.

This builds on our earlier findings that these characteristics are correlated with tracking. We apply the same two-step estimation strategy as specified above, replacing a student's math test score percentile *t* years after 3rd grade with alternative outcomes. From these first-stage regressions, we generate the predicted average class size, peer quality, and likelihood of being above or below grade-level math across the 5 years following grade 3 for each school-cohort at the 25th and 75th percentile of the statewide test score distribution in 3rd grade. We then relate these predicted values to the level of tracking exposure in elementary and middle school.

Table 8 displays our results. With respect to class size, we find that for students at the 25th percentile of the statewide test score distribution in grade 3, exposure to tracking in both elementary and middle school is associated with smaller average class sizes. For students at the and 75th percentile, exposure to tracking in both elementary and middle school is only weakly associated with smaller class sizes. Thus, tracking may benefit lower-achieving students to the extent that it may be accompanied by smaller class sizes.

When we consider the relationship between tracking and average peer achievement, we see that regardless of how we measure peer achievement—either with peers' average test scores in the year prior or peers' average first-observed test scores—for students at the 25th percentile of the statewide test score distribution in grade 3, exposure to tracking is associated with a reduction in peer achievement. In contrast, we find that exposure to tracking is associated with higher achieving peers for students at the 75th percentile. This pattern is to be expected, and is also one possible mechanism through which tracking impacts student achievement. Indeed, peer effects may be part of the explanation for the positive relationship between tracking and the performance of high-achieving students.

Finally, when we examine the likelihood that students are over or under grade-level math, we see that exposure to tracking is associated with a lower likelihood of being above grade level in math for students at the 25th percentile, but a higher likelihood for students at the 75th percentile. Consistent with our earlier findings, these results suggest that tracking may impact achievement by allowing for targeted curricular differentiation based on students' academic performance. The association is stronger for exposure to middle school tracking than exposure to elementary school tracking, which is consistent with the idea that math courses are less differentiated in elementary school. For both lower- and higher-achieving students, we do not find a strong relationship between the likelihood of being below grade-level math and exposure to tracking in either elementary or middle school.

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 and greater curricular differentiation based on students' academic performance. That tracking does not harm lower-achieving students may arise from its association with smaller class sizes and less advanced curriculum.

6. Conclusion

Very little is known about the nature and scope of tracking in the US. In this paper, we use detailed administrative data from Texas to create several 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 much tracking there is across schools and by grade, how schools operationalize tracking, which schools are more likely to track, and how tracking is related to student performance.

We find substantial variation tracking. In addition, in contrast to the popular perception, we find that the amount of ability sorting that takes place within schools is far greater than the amount of ability sorting that occurs across schools. In addition, we find that within-school sorting based on prior test scores is far greater than within-school sorting based on race/ethnicity and SES. We also find that even though within-school math tracking increases as students move from elementary to middle school, math tracking in elementary school translates into a greater degree of sorting throughout the school day relative to tracking in middle school.

We also find that tracking appears to be operationalized through more aggressive classification of students as gifted or disabled and increased curricular differentiation. Our also results suggest that heterogeneity in student achievement is the most important predictor of tracking. Interestingly, once we control for the distribution of student achievement, there is no relationship between tracking and the racial composition of the student body. When we examine the relationship between school accountability and tracking, our results suggest that schools adjust their instruction towards *reduced* tracking concurrent with the receipt of a low performance rating.

Finally, when we examine the implications of tracking for future educational outcomes, we find that while exposure to tracking in elementary and middle is not strongly associated with test score growth for students at the bottom of the initial achievement distribution, exposure to tracking in middle school is positively associated with test score growth for students at the top, suggesting that tracking increases inequities in educational outcomes but does not otherwise harm low-achieving students on average.

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), 90-97.

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), 341-381.

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.

Bradbury, A. (2018). The impact of the Phonics Screening Check on grouping by ability: A necessary evil amid the policy storm. *British Educational Research Journal*, *44*, 539-556.

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200-230.

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.* (Paper Number 18848). National Bureau of Economic Research Working Paper Series.

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), 400-405.

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.

Domina, T., Hanselman, P., Hwang, N., & McEachin, A. (2016). Detracking and tracking up: mathematics course placements in California middle schools, 2003–2013. *American Educational Research Journal*, *53*(4), 1229-1266.

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), 1348–1380.

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., & Deming, R. (2016). Accountability in U.S. higher education: some lessons and design principles. *Journal of Economic Perspectives*, 30(3), 33-56.

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), 923-979.

Goodman, J. (2019). The Labor of Division: Returns to Compulsory High School Math Coursework. *Journal of Labor Economics* 37(4), 1141-1182.

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–C76.

Hashim, S. A., Kane, T.J., Kelley-Kemple, T., Laski, M.E., & Staiger, D.O. 2020. *Have Income-Based Achievement Gaps Widened or Narrowed?* (Paper Number 27714). National Bureau of Economic Research Working Paper Series.

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.

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

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), 169-191.

Macartney, H., & Singleton, J. D. (2018). School boards and student segregation. *Journal of Public Economics* 164, 165-182.

Neal, D., & Schanzenbach, D.W. (2010). Left behind by design: proficiency counts and test-based accountability. *Review of Economics and Statistics*, *92*(2), 263–83.

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.

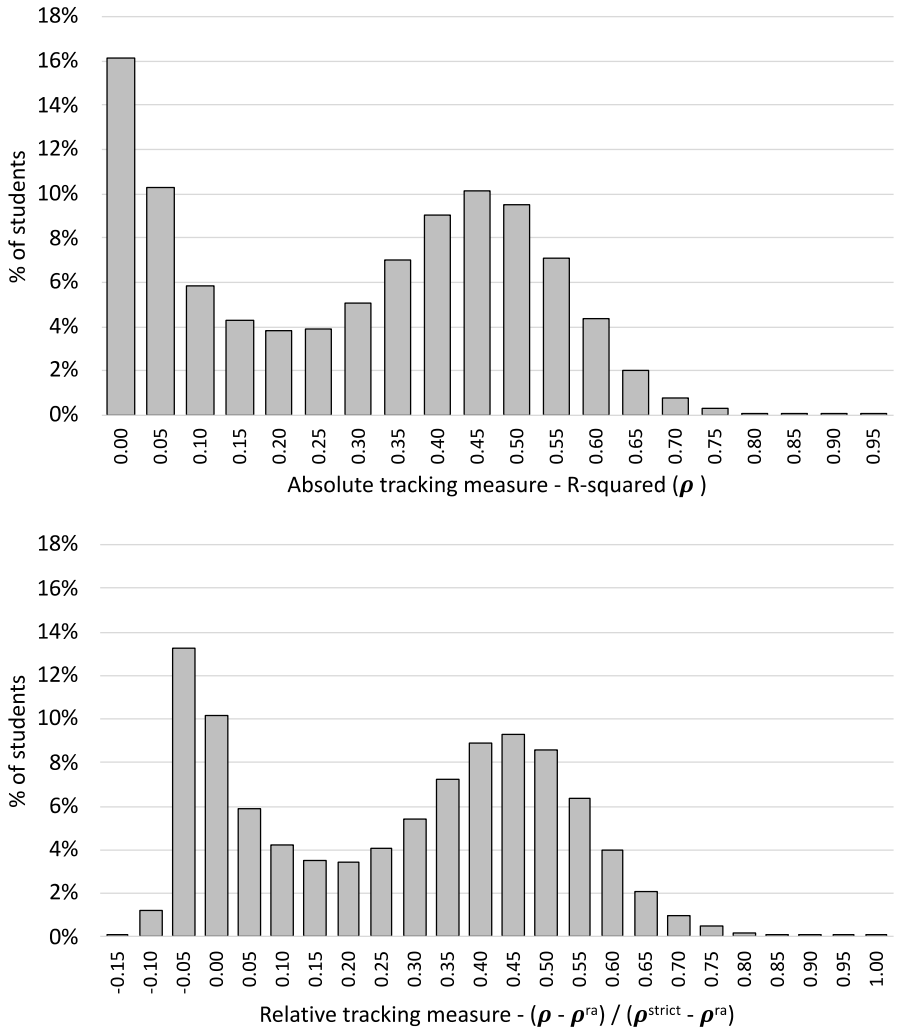
Reback, R. (2008). Teaching to the rating: school accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5-6), 1394-1415.

Reback, R., Rockoff, J., & Schwartz, H.L. (2014). Under pressure: job security, resource allocation, and productivity in schools under no child left behind. *American Economic Journal: Economic Policy*, *6*(3), 207-241.

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.

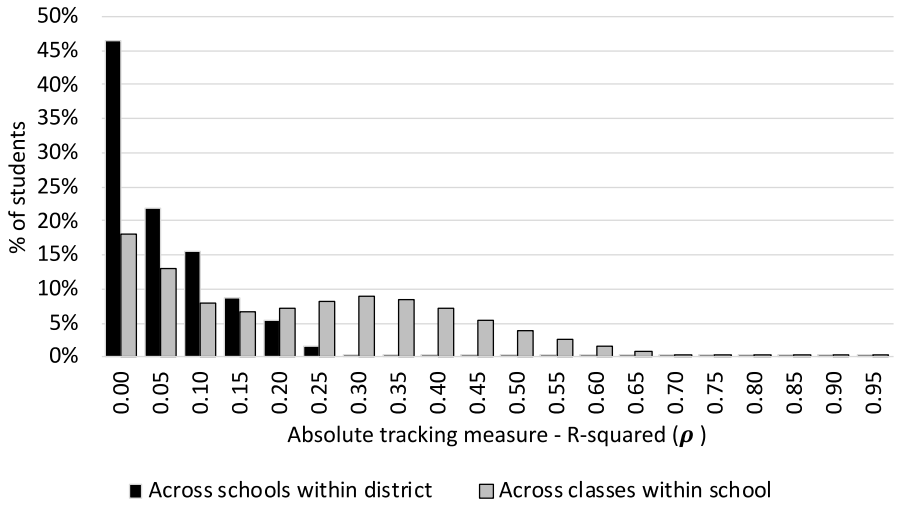
Texas Education Agency. (2020). *Comprehensive biennial report on Texas public schools*. Austin, TX. December, p.179, Table 7.2. https://tea.texas.gov/reports-and-data/school-performance/accountability-research/comprehensive-report-on-texas-public-schools

Figure 1. Distribution of Tracking



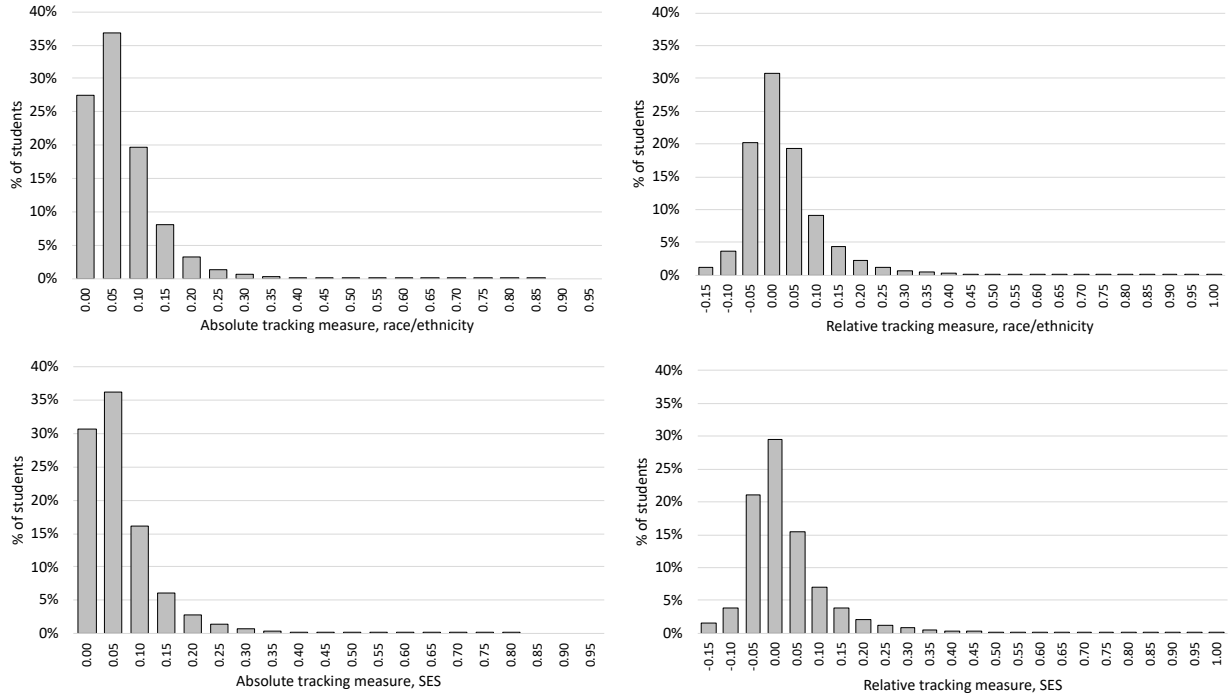
Notes: The top and bottom panels show the student-weighted distributions of the absolute and relative tracking measures, respectively, for the full sample of school-grade-years.

Figure 2. Tracking Within and Across Schools



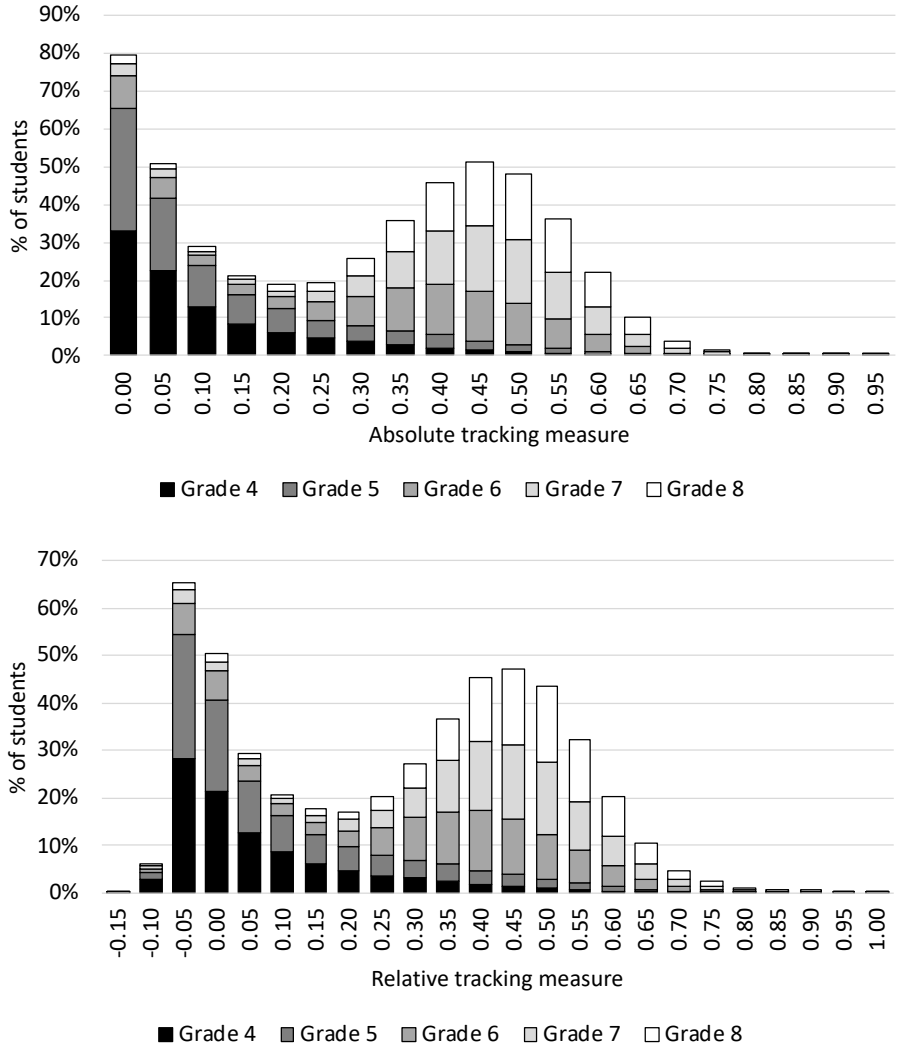
Notes: The figure shows the student-weighted distributions of the absolute tracking measure when tracking is defined as a) the amount of tracking across schools within a district (black bars) and b) our standard measure of tracking across classes within a school (grey bars). For a), the sample includes district-grade-year cells with more than one school.

Figure 3. Distribution of Tracking, by Race/Ethnicity and Socioeconomic Status



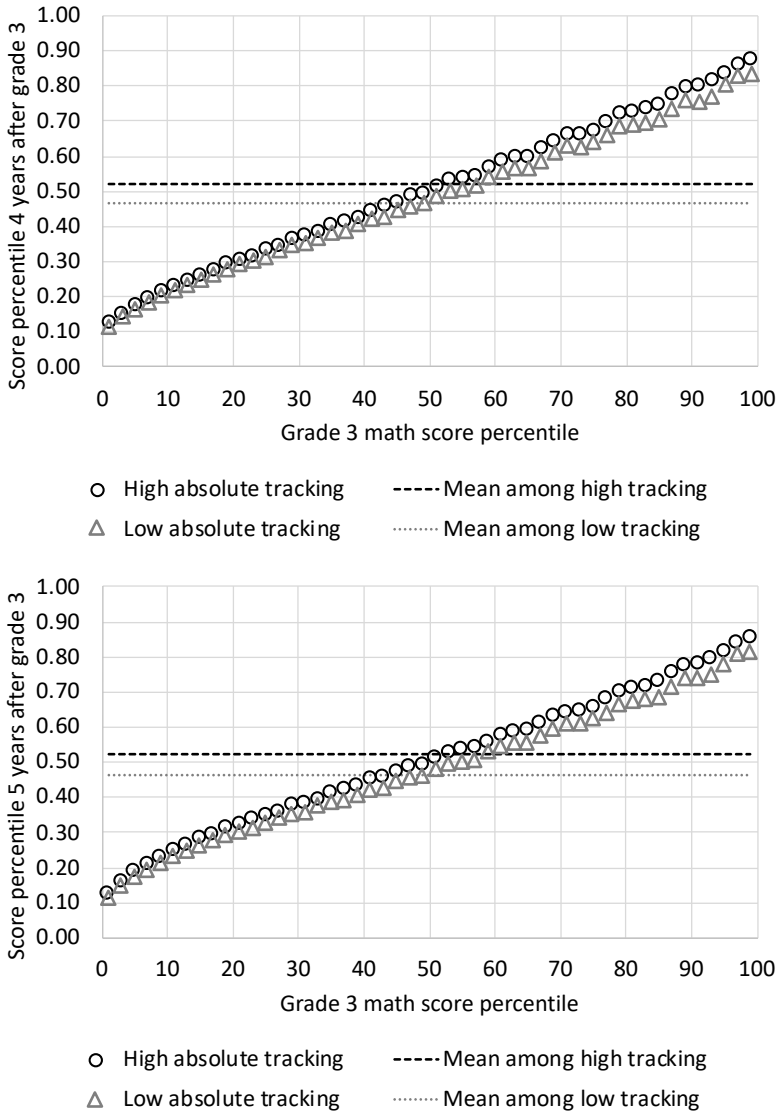
Notes: The panels show the student-weighted distributions of the absolute and relative tracking measures where tracking is measured as tracking across classes by race/ethnicity (defined as Black or Hispanic vs. non-Black and non-Hispanic) in the top panels and by SES (defined as low vs. non-low income) in the bottom panels.

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. Test Score 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-year cohorts with above vs. below average exposure to absolute tracking across those years.

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)
Bilingual two-way	0.009	0.019	0.015	0.005	0.004	0.003
	(0.054)	(0.082)	(0.072)	(0.034)	(0.026)	(0.021)
Bilingual non-two-way	0.055	0.143	0.118	0.015	0.001	0.000
	(0.134)	(0.194)	(0.172)	(0.055)	(0.009)	(0.006)
ESL content-based	0.040	0.036	0.031	0.052	0.044	0.037
	(0.089)	(0.063)	(0.062)	(0.115)	(0.102)	(0.088)
ESL pullout	0.048	0.021	0.021	0.069	0.071	0.060
	(0.083)	(0.050)	(0.049)	(0.103)	(0.096)	(0.085)
Physical disability	0.013	0.012	0.013	0.013	0.012	0.012
	(0.012)	(0.015)	(0.015)	(0.011)	(0.010)	(0.010)
Other disability	0.083	0.081	0.085	0.086	0.083	0.081
	(0.033)	(0.037)	(0.037)	(0.032)	(0.030)	(0.030)
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)
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)
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)
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 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 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. Physical disabilities include disabilities such as orthopedic impairments, auditory and visual impairment, and traumatic brain injuries, while most other disabilities are emotional and learning disabilities. 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. Correlations Between Tracking by Prior Math Scores across Classes, by Subject

		Math		ELA		Science		Social Studies	
		Absolute	Relative	Absolute	Relative	Absolute	Relative	Absolute	Relative
All Grades	Math	1.00	0.99	0.83	0.80	0.76	0.74	0.71	0.69
		0.99	1.00	0.83	0.81	0.76	0.75	0.71	0.70
	ELA	0.83	0.83	1.00	0.99	0.85	0.83	0.83	0.82
		0.80	0.81	0.99	1.00	0.83	0.84	0.83	0.83
	Science	0.76	0.76	0.85	0.83	1.00	0.99	0.90	0.88
		0.74	0.75	0.83	0.84	0.99	1.00	0.89	0.89
	Social Studies	0.71	0.71	0.83	0.83	0.90	0.89	1.00	0.99
		0.69	0.70	0.82	0.83	0.88	0.89	0.99	1.00
Grades 4-5	Math	1.00	0.99	0.90	0.89	0.88	0.87	0.86	0.85
		0.99	1.00	0.89	0.90	0.88	0.88	0.86	0.86
	ELA	0.90	0.89	1.00	0.99	0.94	0.93	0.94	0.93
		0.89	0.90	0.99	1.00	0.93	0.94	0.93	0.94
	Science	0.88	0.88	0.94	0.93	1.00	0.99	0.96	0.95
		0.87	0.88	0.93	0.94	0.99	1.00	0.95	0.96
	Social Studies	0.86	0.86	0.94	0.93	0.96	0.95	1.00	0.99
		0.85	0.86	0.93	0.94	0.95	0.96	0.99	1.00
Grades 6-8	Math	1.00	0.98	0.68	0.64	0.59	0.57	0.54	0.52
		0.98	1.00	0.66	0.65	0.57	0.57	0.52	0.52
	ELA	0.68	0.66	1.00	0.98	0.73	0.71	0.72	0.70
		0.64	0.65	0.98	1.00	0.70	0.71	0.70	0.71
	Science	0.59	0.57	0.73	0.70	1.00	0.99	0.83	0.81
		0.57	0.57	0.71	0.71	0.99	1.00	0.81	0.82
	Social Studies	0.54	0.52	0.72	0.70	0.83	0.81	1.00	0.98
		0.52	0.52	0.70	0.71	0.81	0.82	0.98	1.00

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.

Table 3. 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	0.25	0.45	0.69	0.21	0.39	0.68
Districts with (min) 1 school	2,801	0.27	0.50	0.90	0.22	0.43	0.90
Districts with 2-5 schools	1,483	0.30	0.51	0.64	0.27	0.46	0.60
Districts with 6+ schools	2,392	0.16	0.30	0.36	0.15	0.26	0.31
Grades 4-5							
All districts	4,866	0.31	0.34	0.52	0.27	0.31	0.51
Districts with (min) 1 school	1,887	0.32	0.38	0.77	0.29	0.34	0.77
Districts with 2-5 schools	1,121	0.35	0.39	0.52	0.33	0.36	0.50
Districts with 6+ schools	1,858	0.27	0.29	0.33	0.23	0.25	0.28
Grades 6-8							
All districts	3,071	0.40	0.52	0.79	0.32	0.45	0.80
Districts with (min) 1 school	1,560	0.41	0.54	0.95	0.33	0.47	0.95
Districts with 2-5 schools	621	0.50	0.60	0.74	0.44	0.55	0.70
Districts with 6+ schools	890	0.18	0.29	0.39	0.16	0.26	0.35

Notes: Each cell in columns 2-7 contains the R-squared from a separate 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-by-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 4. Tracking Policies, Absolute Measure of Tracking

	[1]	[2]	[3]	[4]	[5]
Fraction of students:					
Bilingual two-way	-0.004	-0.011	-0.011	0.001	-0.009
	(0.022)	(0.020)	(0.020)	(0.013)	(0.026)
Bilingual non-two-way	0.010	-0.007	-0.007	-0.021	0.000
	(0.022)	(0.021)	(0.021)	(0.013)	(0.016)
ESL content-based	-0.012	-0.011	-0.011	-0.003	-0.028
	(0.026)	(0.023)	(0.023)	(0.037)	(0.024)
ESL pull-out	0.051*	0.064**	0.064**	0.080***	0.034
	(0.031)	(0.029)	(0.029)	(0.023)	(0.029)
Physical disability	0.444***	0.241***	0.240***	0.177***	0.047
	(0.078)	(0.065)	(0.065)	(0.046)	(0.038)
Other disability	0.346***	0.196***	0.199***	0.214***	0.094***
	(0.057)	(0.052)	(0.052)	(0.027)	(0.020)
Gifted	0.205***	0.149***	0.147***	0.151***	0.076***
	(0.042)	(0.040)	(0.041)	(0.033)	(0.022)
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)
Curricular differentiation		0.199***	0.197***	0.263***	0.251***
		(0.026)	(0.026)	(0.031)	(0.031)
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: Each observation is a school-grade-year cell. Observations are weighted by cell enrollment. In addition to the coefficients displayed in the table, all specifications contain the following controls: grade and year indicators, log of school-grade-year enrollment, the mean and standard deviation of prior math test scores in the school-grade-year cell, 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 previous scores within the school-grade-year (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.

Table 5. 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)	
County Democratic vote share			-0.008	-0.007	n/a	n/a
			(0.032)	(0.032)		
District private school share			-0.253*	-0.258*	n/a	n/a
			(0.133)	(0.133)		
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			-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: Each observation is a school-grade-year cell. Observations are weighted by cell enrollment. In addition to the coefficients displayed in the table, all specifications contain the following controls: grade and year indicators, log of school-grade-year enrollment, the mean and standard deviation of prior math test scores in the school-grade-year cell, 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 previous scores within the school-grade-year (i.e., 10th, 25th, 75th, and 90th percentiles) and fixed effects at the district-by-grade or school-by-grade level. "County Democratic vote share" is the average two-party Democratic vote share across the 2000-2016 presidential elections. "District private school share" is the 2010-2016 average share of families with children enrolled in private school. Standard errors are clustered by district, and charter schools are assigned to the geographic districts within which they reside.

Table 6. Effects of Low Accountability Ratings on Tracking

	Absolute measure [1]	Relative measure [2]
Year T-2 (excluded)		
Year T-1	-0.009	-0.008
	(0.006)	(0.008)
Year T	-0.029***	-0.037***
	(0.007)	(0.007)
Year T+1	-0.027***	-0.034***
	(0.007)	(0.008)
Year T+2	-0.017**	-0.024***
	(0.008)	(0.009)

Notes: Each observation is a school-grade-year cell. Observations are weighted by cell enrollment. We employ an event-study, so that each coefficient estimates the within school-grade difference between tracking in the year in question relative to tracking in the excluded year (two years before the school received a low accountability rating). We include school-by-grade fixed effects and year dummies, as well as: log of school-grade-year 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, indicators for whether the district is classified as suburban, town, or rural (so that urban districts are the omitted case), and a set of student demographic controls (fraction Black, Hispanic, Asian, other race/ethnicity, limited English proficient, and low income). We exclude schools that received a low accountability rating in 2011 or earlier or 2018 or later, and we exclude school-by-grade cells that are ever missing observations during the sample period. Standard errors are clustered by district, and charter schools are assigned to the geographic districts within which they reside.

Table 7. Effects of Tracking on Achievement Mobility

	Absolute measure				Relative 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	
Independent variable	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	-0.029**	-0.041*	-0.019	-0.041	-0.024***	-0.037**	-0.020	-0.035
	(0.012)	(0.022)	(0.018)	(0.030)	(0.009)	(0.018)	(0.016)	(0.028)
Middle school tracking	-0.014	0.010	0.019	0.043	-0.009	0.007	0.017	0.035
	(0.016)	(0.023)	(0.018)	(0.032)	(0.013)	(0.017)	(0.015)	(0.024)
	Grade 3 + 3				Grade 3 + 3			
Elementary school tracking	-0.012	-0.006	0.004	0.012	-0.010	-0.008	0.002	0.007
	(0.010)	(0.019)	(0.010)	(0.021)	(0.008)	(0.015)	(0.008)	(0.017)
Middle school tracking	-0.011	-0.004	0.050***	0.050*	-0.009	-0.001	0.044***	0.044*
	(0.017)	(0.026)	(0.019)	(0.029)	(0.014)	(0.020)	(0.015)	(0.023)
	Grade 3 + 4				Grade 3 + 4			
Elementary school tracking	-0.006	-0.026	0.009	0.003	-0.004	-0.021	0.006	0.000
	(0.009)	(0.017)	(0.008)	(0.018)	(0.007)	(0.013)	(0.007)	(0.014)
Middle school tracking	0.009	0.027	0.083***	0.105***	0.007	0.026*	0.071***	0.089***
	(0.012)	(0.019)	(0.018)	(0.026)	(0.010)	(0.016)	(0.015)	(0.021)
	Grade 3 + 5				Grade 3 + 5			
Elementary school tracking	-0.011	-0.016	0.009	0.002	-0.008	-0.010	0.007	0.005
	(0.009)	(0.020)	(0.008)	(0.019)	(0.007)	(0.015)	(0.007)	(0.016)
Middle school tracking	0.009	0.019	0.095***	0.132***	0.013	0.023	0.087***	0.120***
	(0.017)	(0.027)	(0.020)	(0.029)	(0.013)	(0.022)	(0.017)	(0.023)

Notes: Each pair of estimated coefficients on elementary and middle school tracking comes from a separate school-cohort level regression. 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 (75th) percentile of the statewide distribution. 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; for calculating these instruments, we also restrict attention to students who have enrollment records for each grade 4-8 and who do not repeat any grades during that period. All regressions include as controls a set of school and cohort fixed effects, the mean and standard deviation of grade 3 math scores in the school-cohort, and for each year after grade 3, the fraction of the school-cohort with enrollment records in that year. Standard errors are clustered by district.

Table 8. Effects of Tracking on Educational Inputs

	Absolute measure				Relative measure			
	Predicted outcome for students at 25th percentile in grade 3		Predicted outcome for students at 75th percentile in grade 3		Predicted outcome for students at 25th percentile in grade 3		Predicted outcome for students at 75th percentile in grade 3	
Independent variable	OLS [1]	IV [2]	OLS [3]	IV [4]	OLS [5]	IV [6]	OLS [7]	IV [8]
	Class size				Class size			
Elementary school tracking	-2.827***	-2.967***	-1.571***	-1.749*	-1.096***	-1.426**	-0.163	-0.429
	(0.435)	(0.986)	(0.416)	(0.947)	(0.269)	(0.703)	(0.259)	(0.663)
Middle school tracking	-2.039***	-0.962	-0.828*	0.050	-1.202***	-0.414	-0.302	0.353
	(0.439)	(0.638)	(0.445)	(0.614)	(0.350)	(0.499)	(0.351)	(0.466)
	Peers' initial math z-score				Peers' initial math z-score			
Elementary school tracking	-0.375***	-0.371***	0.387***	0.373***	-0.300***	-0.299***	0.313***	0.306***
	(0.014)	(0.021)	(0.014)	(0.025)	(0.012)	(0.017)	(0.011)	(0.020)
Middle school tracking	-0.180***	-0.152***	0.303***	0.343***	-0.148***	-0.124***	0.264***	0.295***
	(0.016)	(0.022)	(0.019)	(0.021)	(0.014)	(0.018)	(0.016)	(0.018)
	Peers' previous math z-score				Peers' previous math z-score			
Elementary school tracking	-0.383***	-0.364***	0.402***	0.398***	-0.307***	-0.298***	0.326***	0.325***
	(0.018)	(0.030)	(0.019)	(0.035)	(0.014)	(0.024)	(0.015)	(0.028)
Middle school tracking	-0.211***	-0.187***	0.415***	0.439***	-0.177***	-0.155***	0.359***	0.375***
	(0.027)	(0.036)	(0.031)	(0.040)	(0.023)	(0.029)	(0.025)	(0.033)
	Likelihood over grade-level in math in grade 3+5				Likelihood over grade-level in math in grade 3+5			
Elementary school tracking	-0.034***	-0.066**	0.008	-0.057	-0.026**	-0.054**	0.006	-0.041
	(0.013)	(0.030)	(0.023)	(0.049)	(0.010)	(0.025)	(0.018)	(0.039)
Middle school tracking	-0.118***	-0.080	0.334***	0.458***	-0.088***	-0.052	0.314***	0.423***
	(0.039)	(0.052)	(0.061)	(0.093)	(0.033)	(0.042)	(0.051)	(0.074)
	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	0.013	0.000	0.001	-0.001	0.009	0.000	0.000
	(0.006)	(0.012)	(0.002)	(0.003)	(0.005)	(0.010)	(0.001)	(0.003)
Middle school tracking	-0.033***	0.001	0.003	-0.003	-0.027***	0.003	0.002	-0.003
	(0.010)	(0.015)	(0.004)	(0.006)	(0.009)	(0.013)	(0.003)	(0.005)

Notes: Each pair of estimated coefficients on elementary and middle school tracking comes from a separate school-cohort level regression. The outcomes in the top three panels are the predicted average class size and peer quality across the 5 years following grade 3 for students from the school-cohort at the 25th and 75th percentile of the statewide test score distribution in grade 3. The outcomes in the bottom 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. 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; for calculating these instruments, we also restrict attention to students who have enrollment records for each grade 4-8 and who do not repeat any grades during that period. All regressions include as controls a set of school and cohort fixed effects, the mean and standard deviation of grade 3 math scores in the school-cohort, and for each year after grade 3, the fraction of the school-cohort with enrollment records in that year. Standard errors are clustered by district.

ONLINE APPENDICES

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.

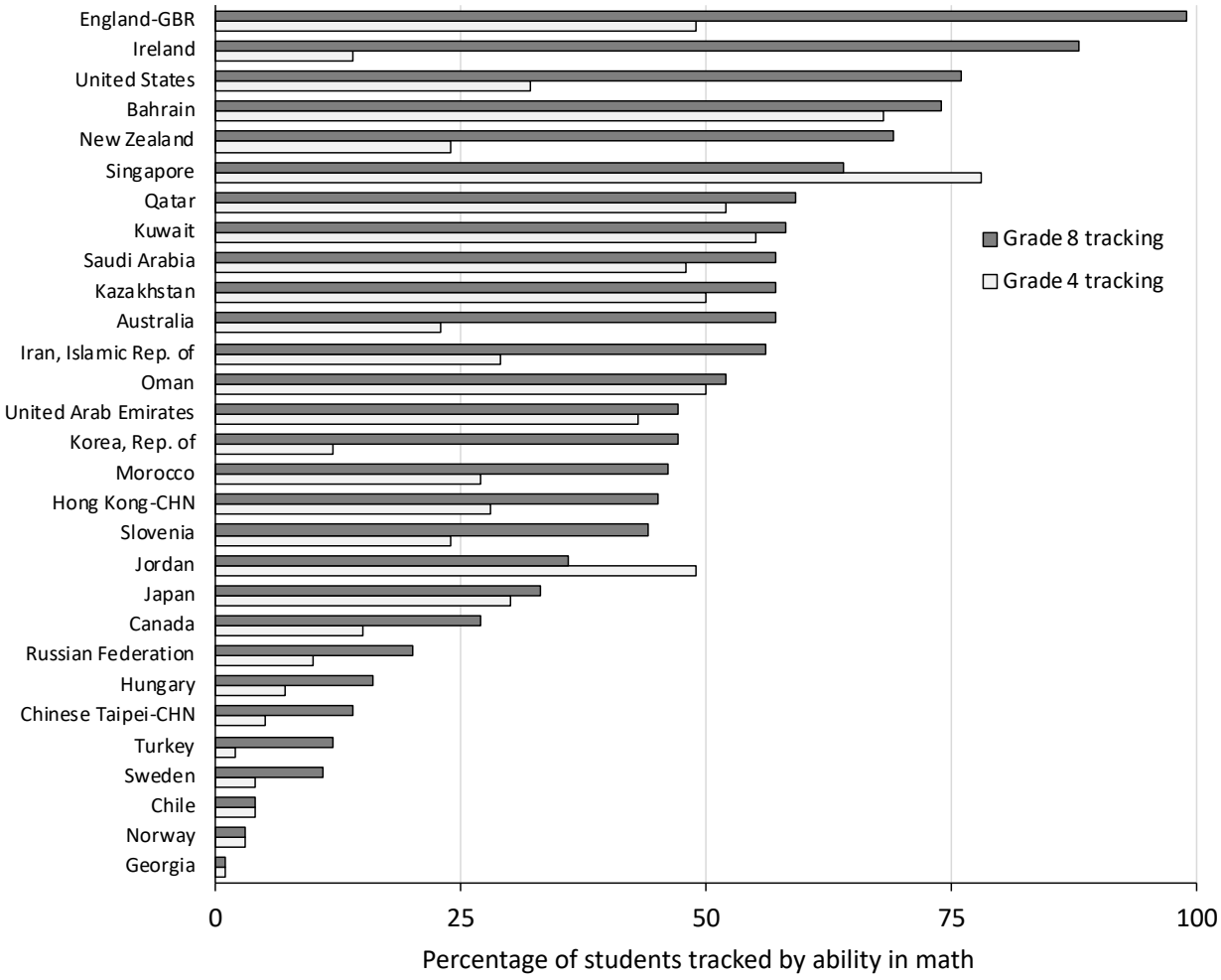
Figure A1 places the US experience in the context of other countries. It reports statistics from the 2015 Trends in International Mathematics and Science Study (TIMMS) for rates of within-school tracking in 4th and 8th grade by participating country. Regardless of the grade, the US exhibits high rates 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.

References

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).

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).

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 and their properties, as well as 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 $A = \{a_1, a_2, \dots\}$ be the set of students in a school-grade-year cohort, let $C = \{c_1, c_2, \dots\}$ 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:

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

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:[36]

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

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:[37]

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

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.

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

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

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:

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

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 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.[38]

[38] 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.

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 of tracking 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' 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 bounded above, restricted in value 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) have the smallest values.

With these two statistics, we develop an alternative measure of tracking that accounts for differences in the class size and achievement distributions across cohorts. We interpret the random assignment counterfactual as a lack of any tracking policy, and we interpret the purposeful assignment counterfactual as the most intense tracking policy possible. Therefore, we define:

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

This measure of relative tracking 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.

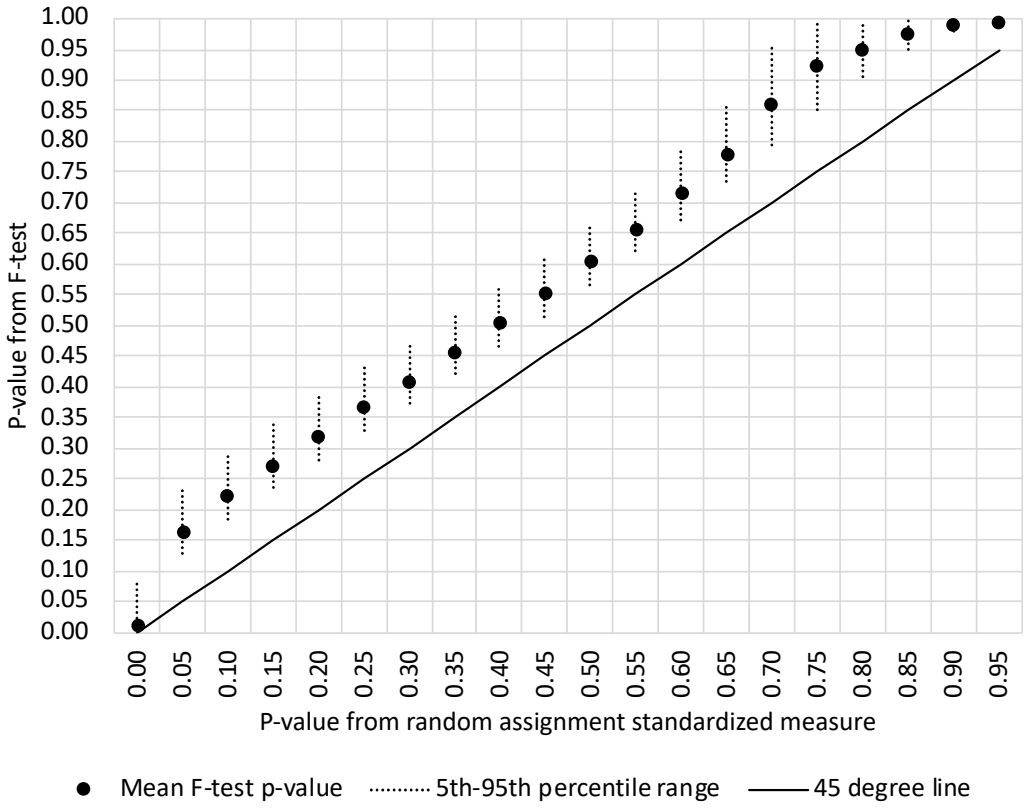
This measure ρ^{rel} is related to the "effective network isolation index" in Hellerstein et al. (2011). They standardize their index of network isolation (in the context of racial segregation) using the mean of that index from simulations with random assignment as well as the maximum value the index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

References

Collins, Courtney A. and Li Gan. 2013. Does Sorting Students Improve Scores? An Analysis of Class Composition. NBER Working Paper Number 18848.

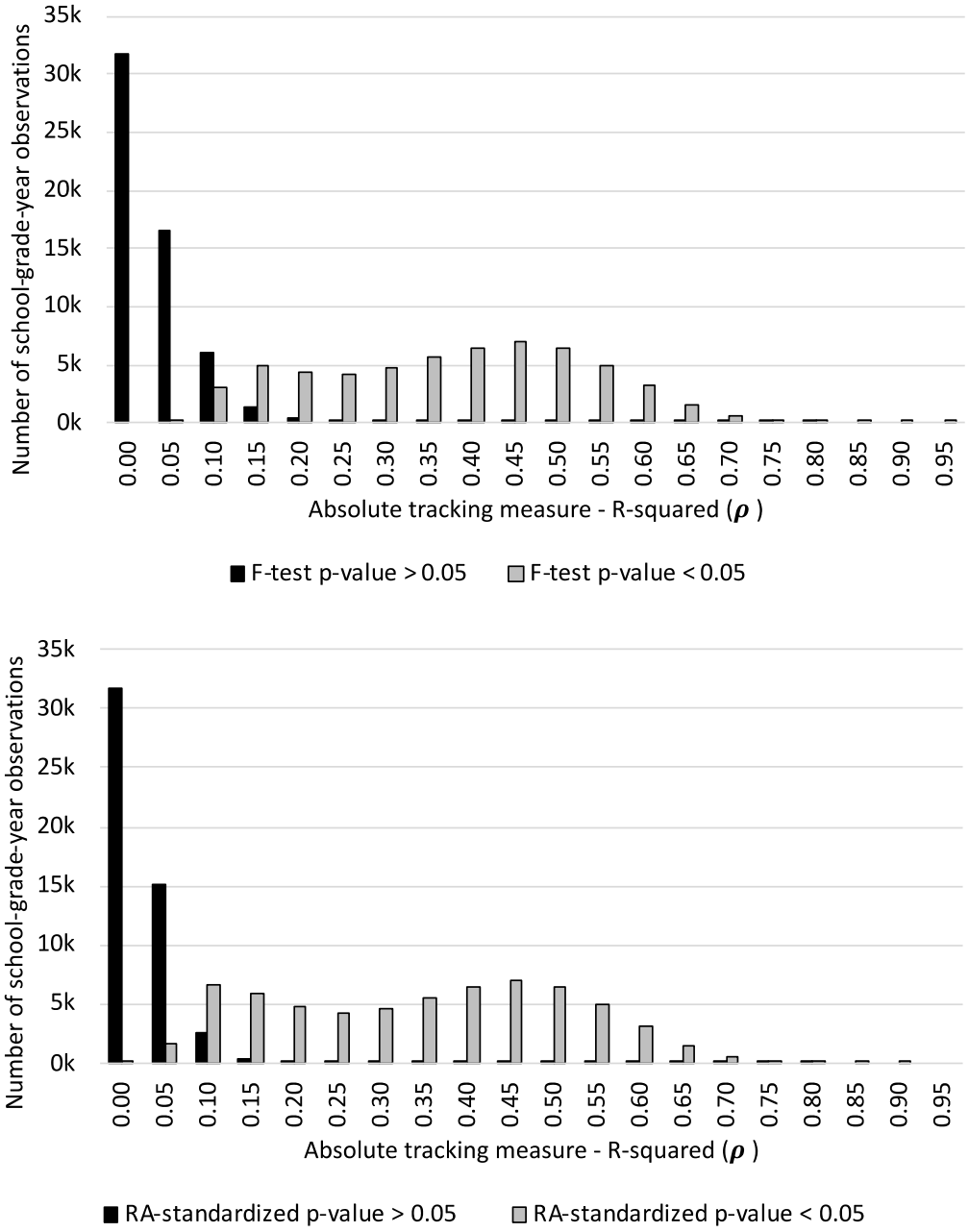
Hellerstein, J. K., McInerney, M., & Neumark, D. (2011). Neighbors and Coworkers: The Importance of Residential Labor Market Networks. *Journal of Labor Economics*, 29(4), 659–695. https://doi.org/10.1086/660776

Figure B1. Comparison of P-values across Approaches



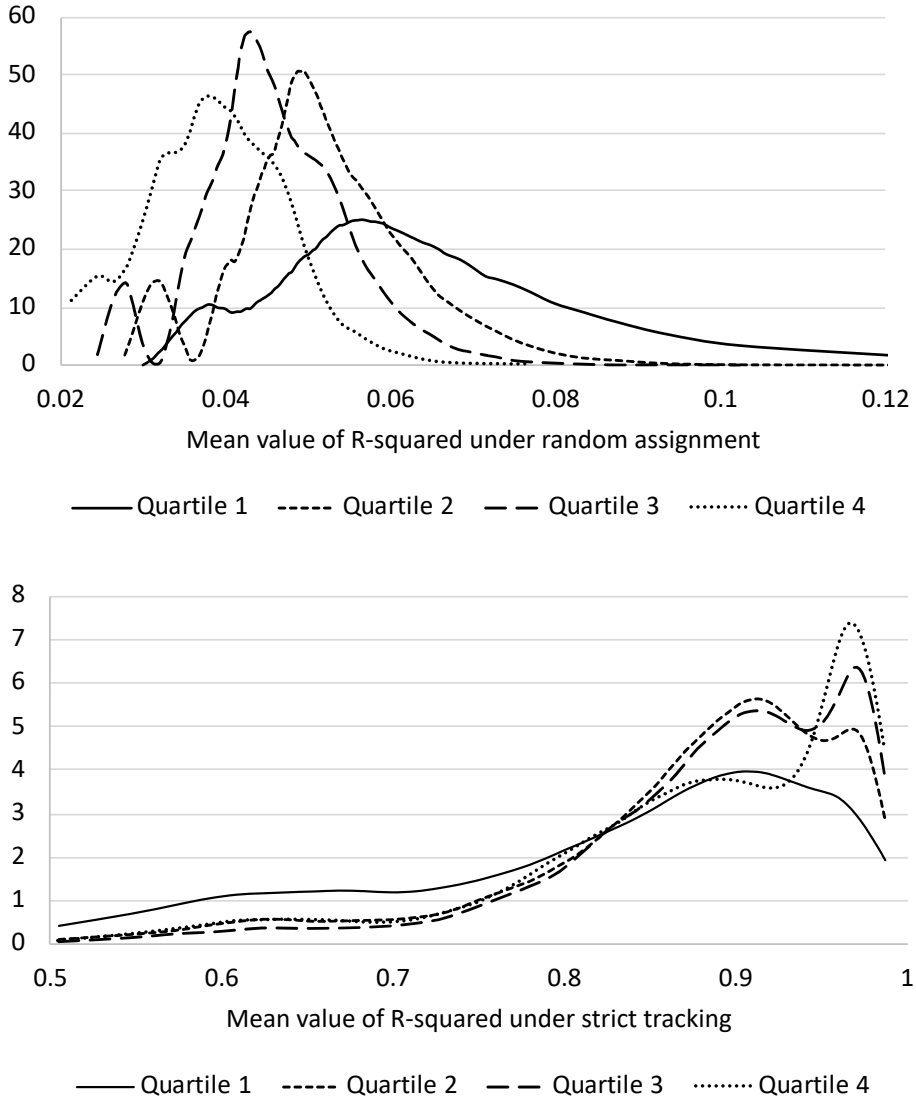
Notes: This figure compares the p-values from the F-test of the joint significance of the class indicators in the regression predicting prior achievement with those from the finite sample approach based on random assignment of students to classes. On the x-axis, the first bin is 0-0.05, the second bin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

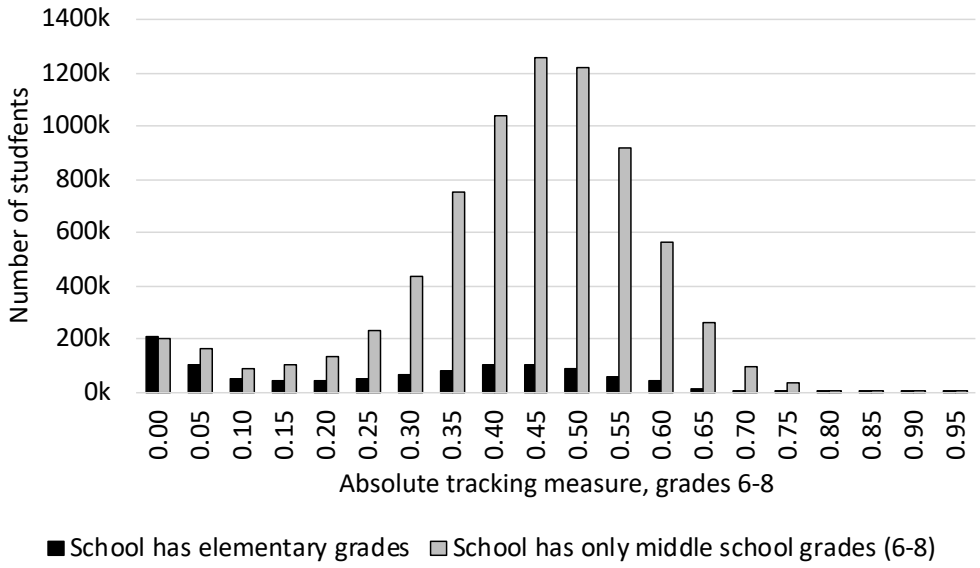
Figure B3. Distribution of the Mean R-squared under Random and Strict Assignment, by Average Class Size



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 12, 16, 19 and 23 students moving from quartile 1 to quartile 4.

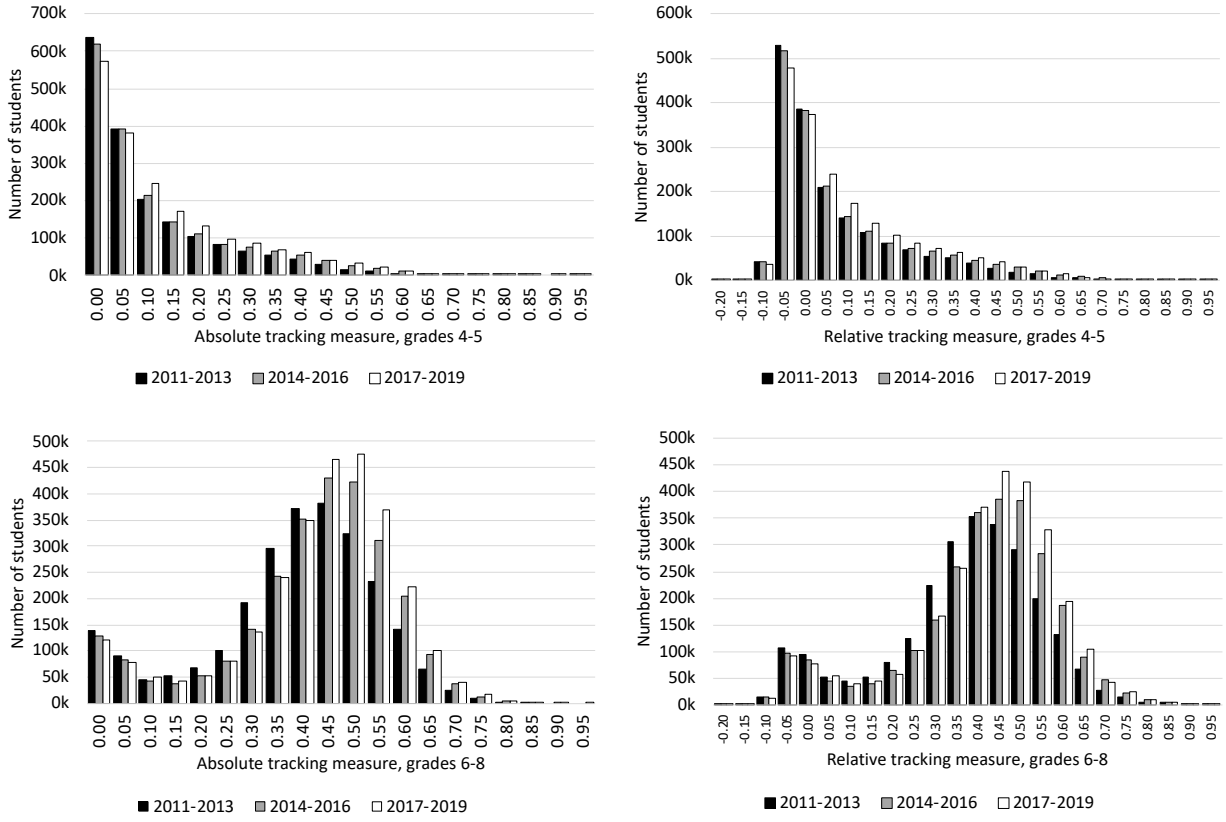
Appendix C. Supplementary Figures and Tables

Figure C1. Absolute Tracking Measure for Grades 6-8, by School Grade Composition



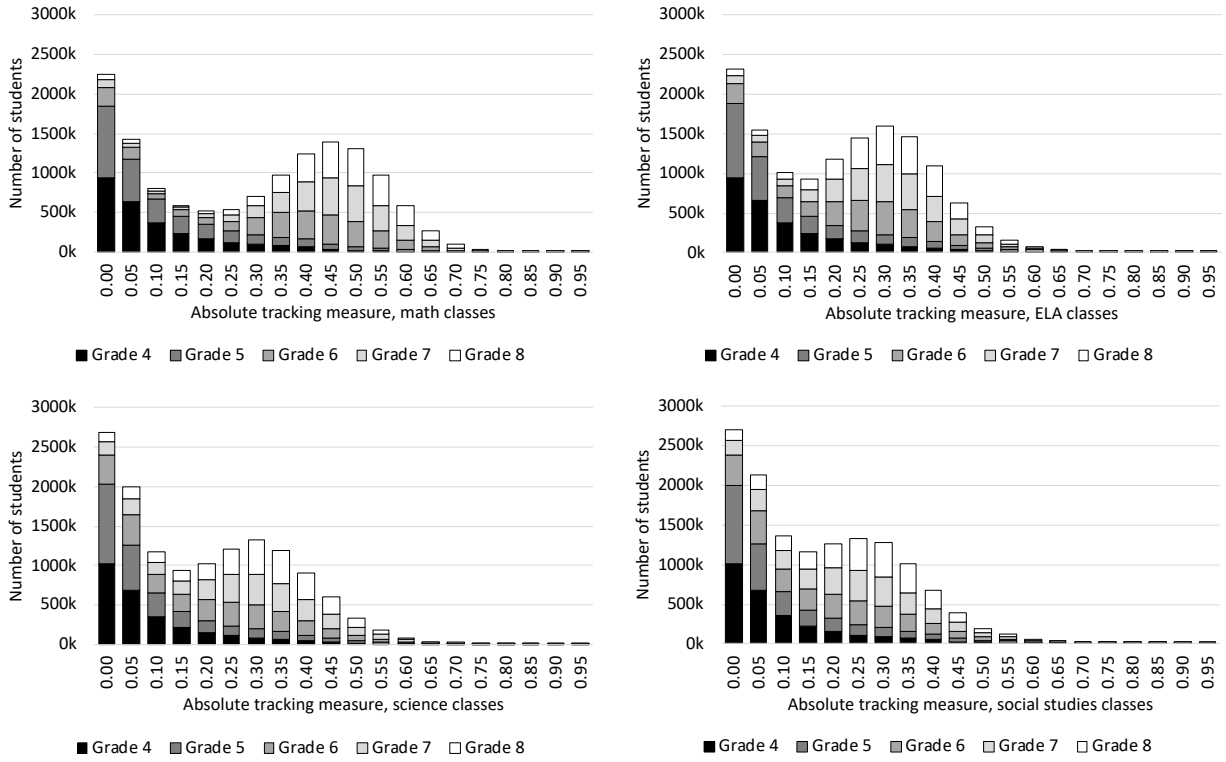
Notes: This figure shows the student-weighted distribution of the absolute tracking measure for students in middle school grades (6-8), broken down by whether the school serves any grades below grade 6.

Figure C2. Tracking over Time



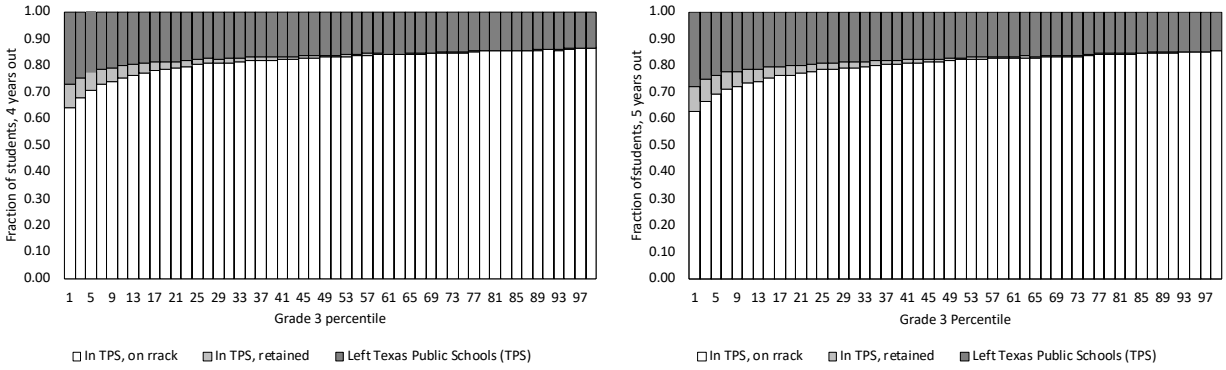
Notes: This figure shows the student-weighted distributions of the absolute and relative tracking measures, broken down by grade-level and time periods.

Figure C3. Absolute Tracking Measures for Math and Other Subjects



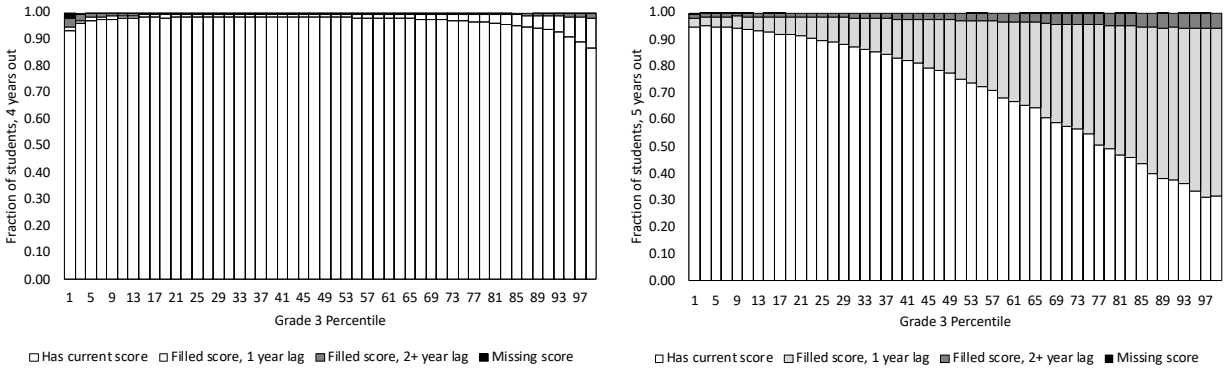
Notes: These panels show the student-weighted distribution of absolute tracking by prior math scores for math (top left), English language arts/reading (top right), science (bottom left), and social studies (bottom right) classes, broken down by grade.

Figure C4. 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 C5. 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.

Table C1. Total Variation in Prior Math Test Scores Accounted for by District/School/Class

	Variance in 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
All students	0.10	0.17	0.44	0.29	0.37	0.43	0.22	0.33	0.39
Districts with (minimum) 1 school	0.14	0.15	0.41	0.33	0.34	0.40	0.21	0.23	0.30
Districts with 2-5 schools	0.10	0.15	0.44	0.30	0.36	0.42	0.25	0.32	0.39
Districts with 6+ schools	0.07	0.19	0.46	0.23	0.37	0.43	0.21	0.39	0.45
Grades 4-5									
All districts	0.08	0.16	0.29	0.29	0.39	0.45	0.22	0.36	0.41
Districts with (minimum) 1 school	0.12	0.14	0.26	0.33	0.35	0.40	0.21	0.24	0.30
Districts with 2-5 schools	0.08	0.16	0.29	0.30	0.38	0.44	0.24	0.35	0.41
Districts with 6+ schools	0.05	0.17	0.30	0.23	0.40	0.45	0.22	0.43	0.48
Grades 6-8									
All districts	0.11	0.18	0.55	0.30	0.36	0.42	0.22	0.31	0.38
Districts with (minimum) 1 school	0.15	0.16	0.50	0.32	0.33	0.39	0.21	0.22	0.30
Districts with 2-5 schools	0.11	0.15	0.54	0.31	0.35	0.41	0.25	0.30	0.38
Districts with 6+ schools	0.08	0.20	0.57	0.23	0.35	0.42	0.21	0.36	0.43

Notes: Districts are grouped by the minimum number of schools for any grade-year across grades 4-8 and years 2011-2019. The R-squared is reported in each cell 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) on a set of indicators for each district, school, or class, as indicated in the column sub-header.