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.

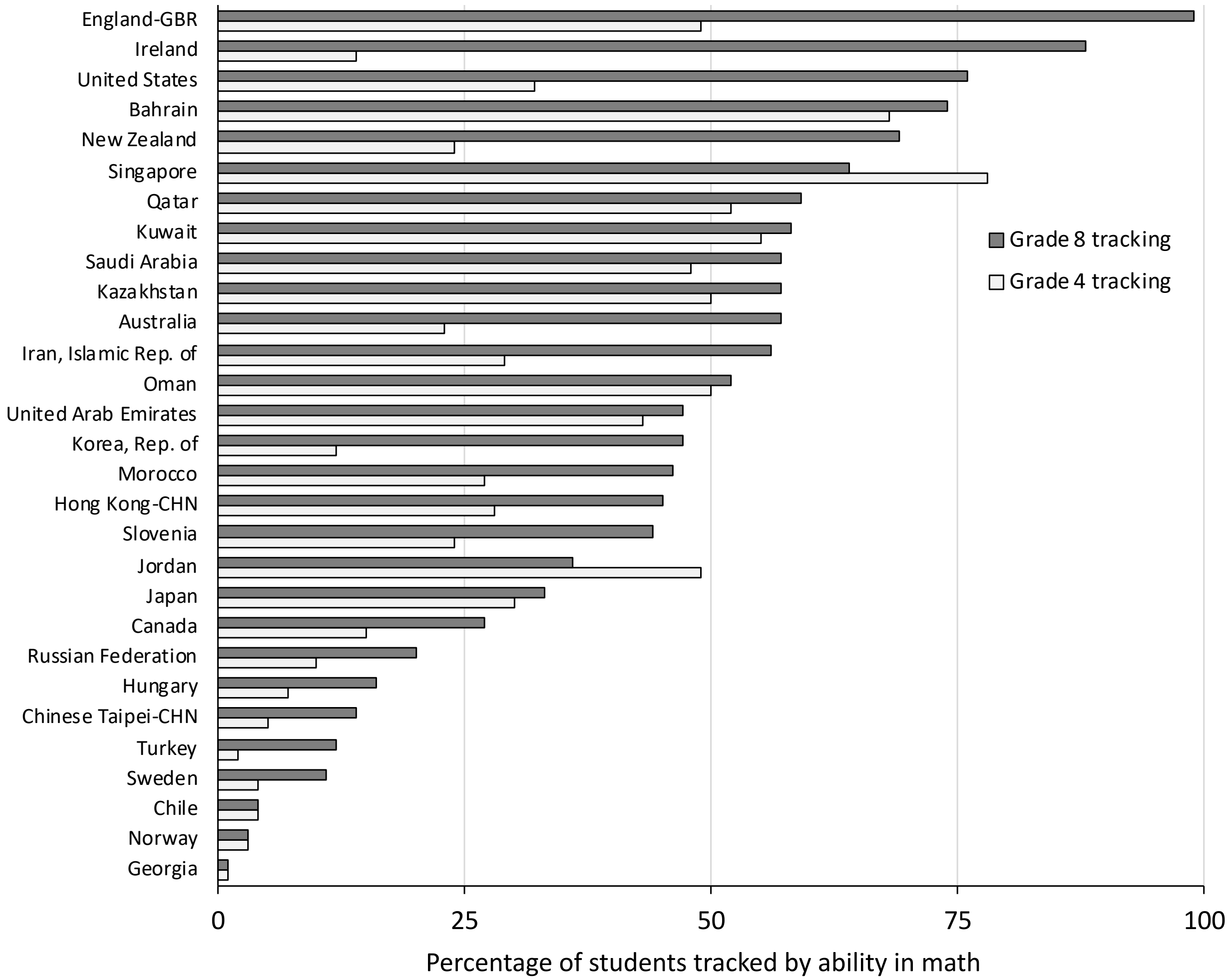
Figure A1 places the US experience in the context of other countries. It reports statistics from the 2015 Trends in International Mathematics and Science Study (TIMMS) for rates of within-school tracking in 4th and 8th grade by participating country. Regardless of the grade, the US exhibits high rates 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 $A = \{a_1, a_2, \ldots\}$ be the set of students in a school-grade-year cohort, let $C = \{c_1, c_2, \ldots\}$ 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:[1]

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

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:[2]

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

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

B.2 Statistical Significance

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

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

$$p^F = 1 - F_{N^C, N-N^C-1}(F)$$

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

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

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

[1] 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.

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

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

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

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

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

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

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

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

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

$$\begin{aligned} E_P(\kappa_p) &= \frac{1}{N}\sum_{c \in C}\frac{1}{N_c}E_P\left(\left(\sum_{a \in b_c} x_a\right)^2\right) = \frac{1}{N}\sum_{c \in C}\frac{1}{N_c}E_P\left(\sum_{a \in b_c} x_a^2 + \sum_{a \in b_c}\sum_{j \neq a} x_a x_j\right) \\ &= \frac{1}{N}\sum_{c \in C}\frac{1}{N_c}\left(N_c E(x_a^2) + N_c(N_c - 1)E\left(x_a x_j | a \neq j\right)\right) \\ &= \frac{N^C}{N}E(x_a^2) + \frac{N-N^C}{N}E\left(x_a x_j | a \neq j\right) = \frac{N^C - 1}{N-1}\eta + \frac{N-N^C}{N-1}\lambda \end{aligned}$$

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

For that reason, rather than simulate $\rho^{ra,\mu}$ and $\rho^{ra,\sigma}$, we calculate these moments.[3]

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

[3] 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.

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.

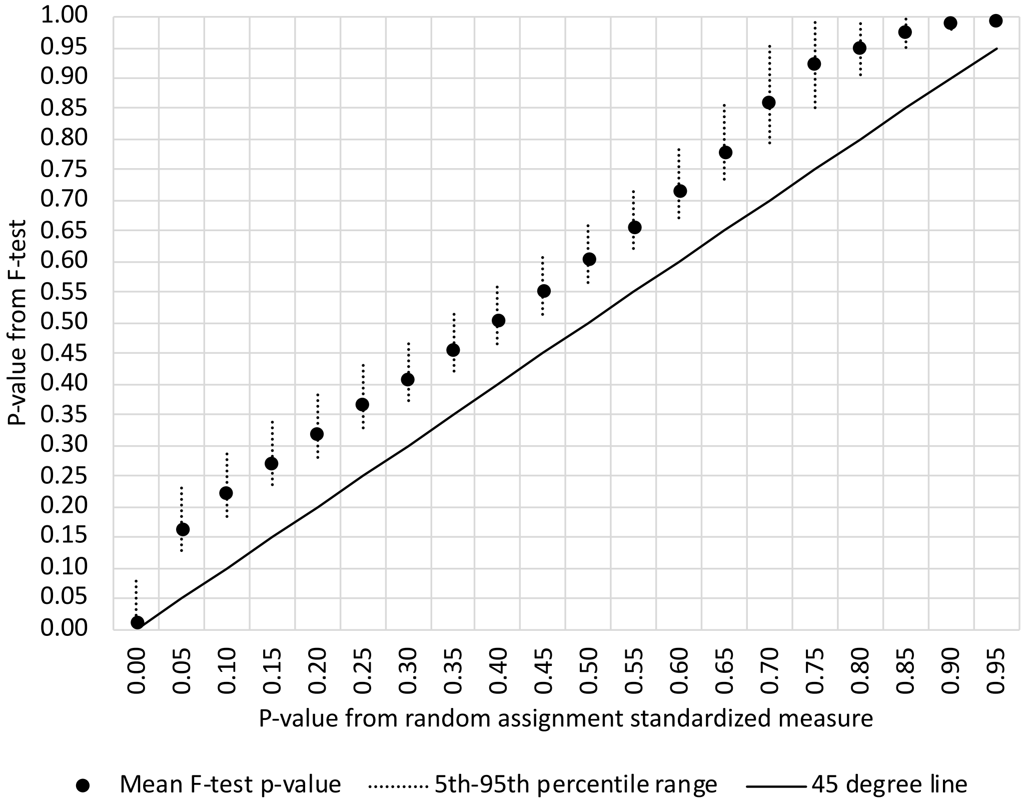
This measure ρ^{rel} is related to the "effective network isolation index" in Hellerstein et al. (2011). They standardize their index of network isolation (in the context of racial segregation) using the mean of that index from simulations with random assignment as well as the maximum value the index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

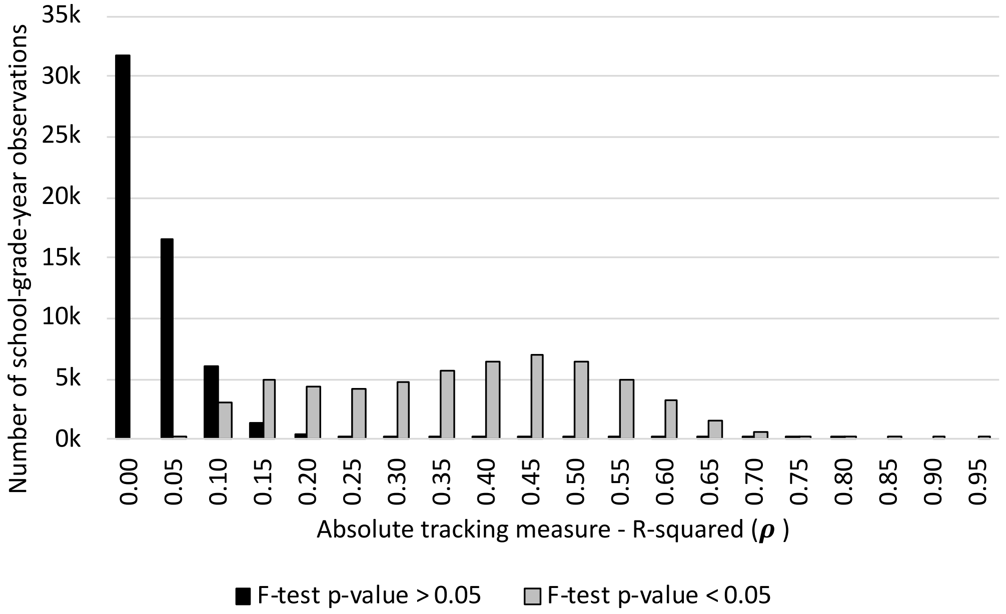
Hellerstein, J. K., McInerney, M., & Neumark, D. (2011). Neighbors and Coworkers: The Importance of Residential Labor Market Networks. *Journal of Labor Economics*, 29(4), 659–695. https://doi.org/10.1086/660776

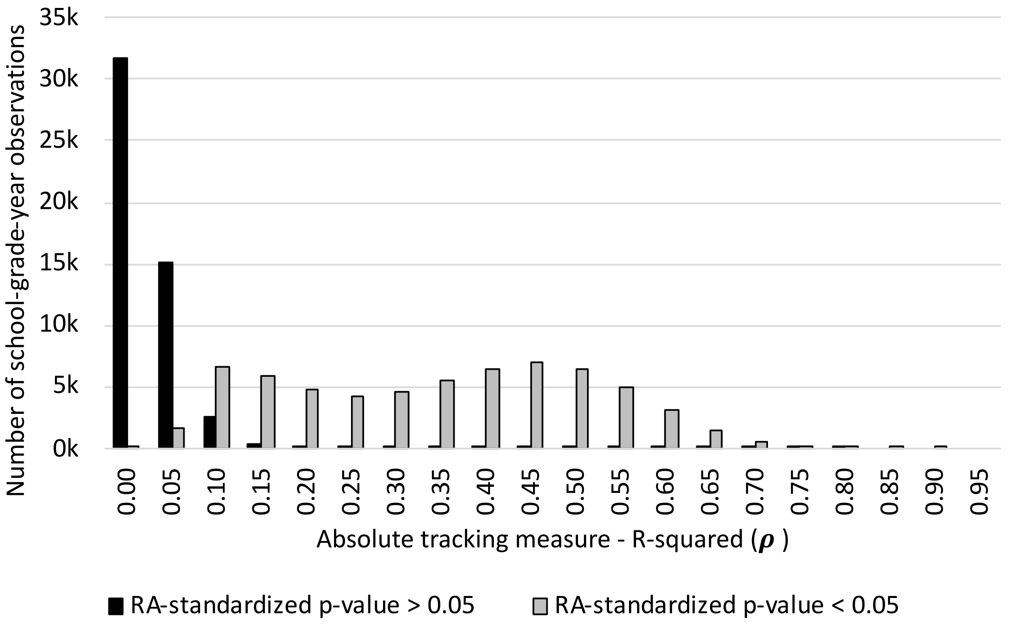
Figure B1. Comparison of P-values across Approaches



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.

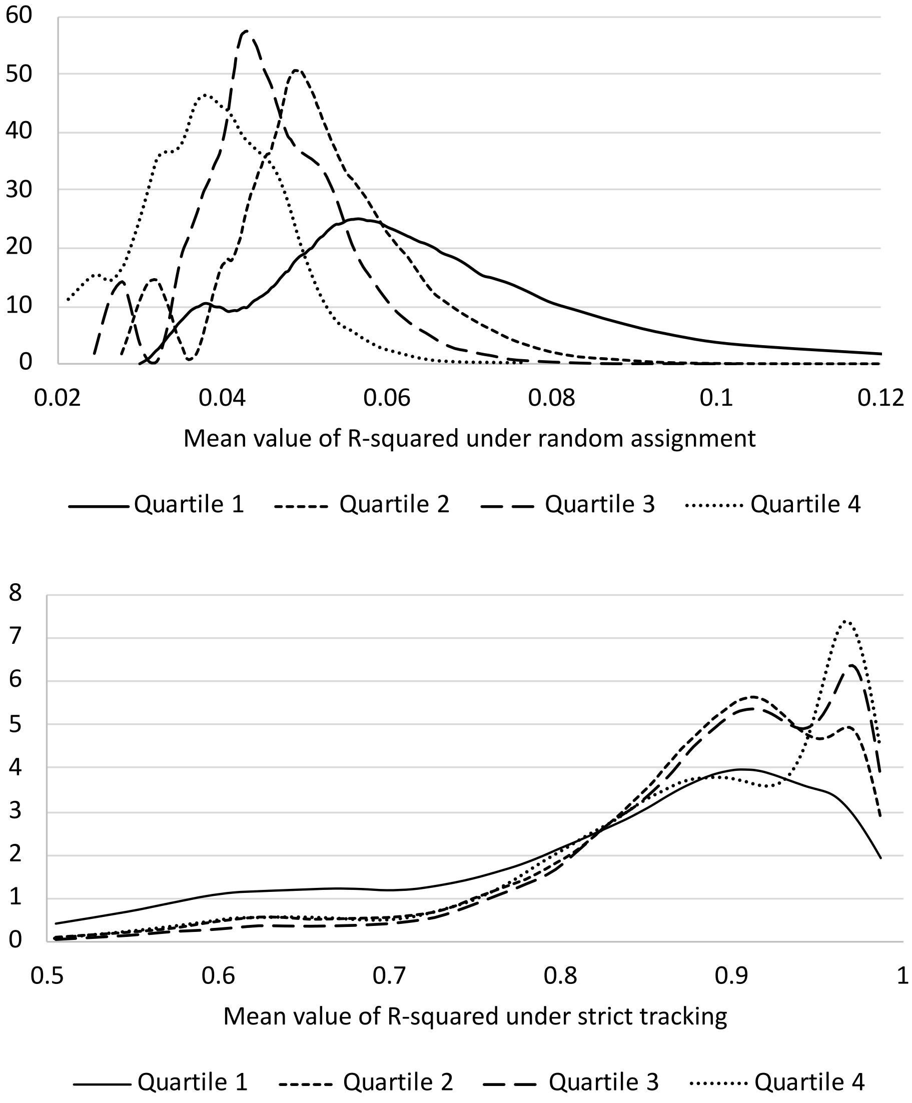
Figure B2. Level of Tracking by Confidence in Tracking, by Approach





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.

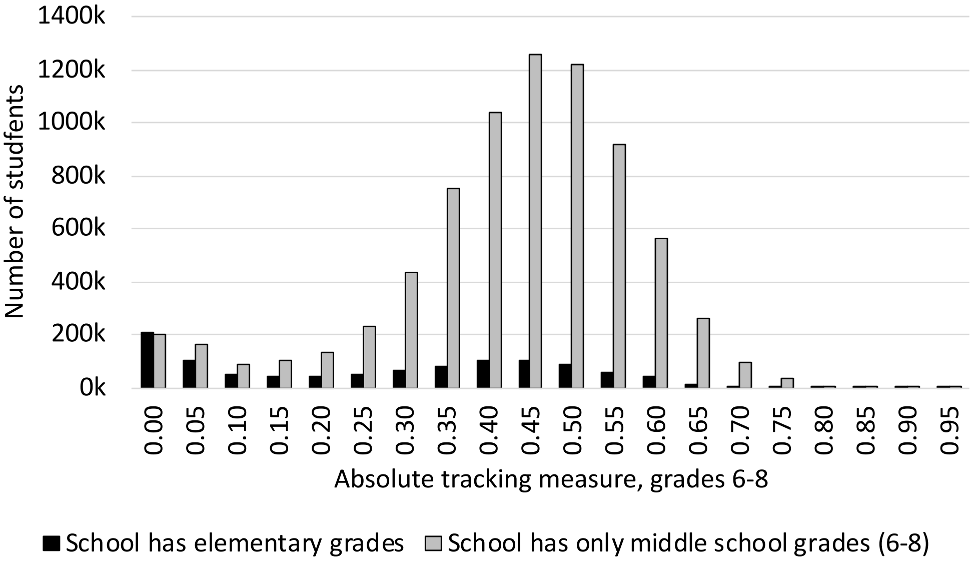
Figure B3. Distribution of the Mean R-squared under Random and Strict Assignment, by Average Class Size



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.

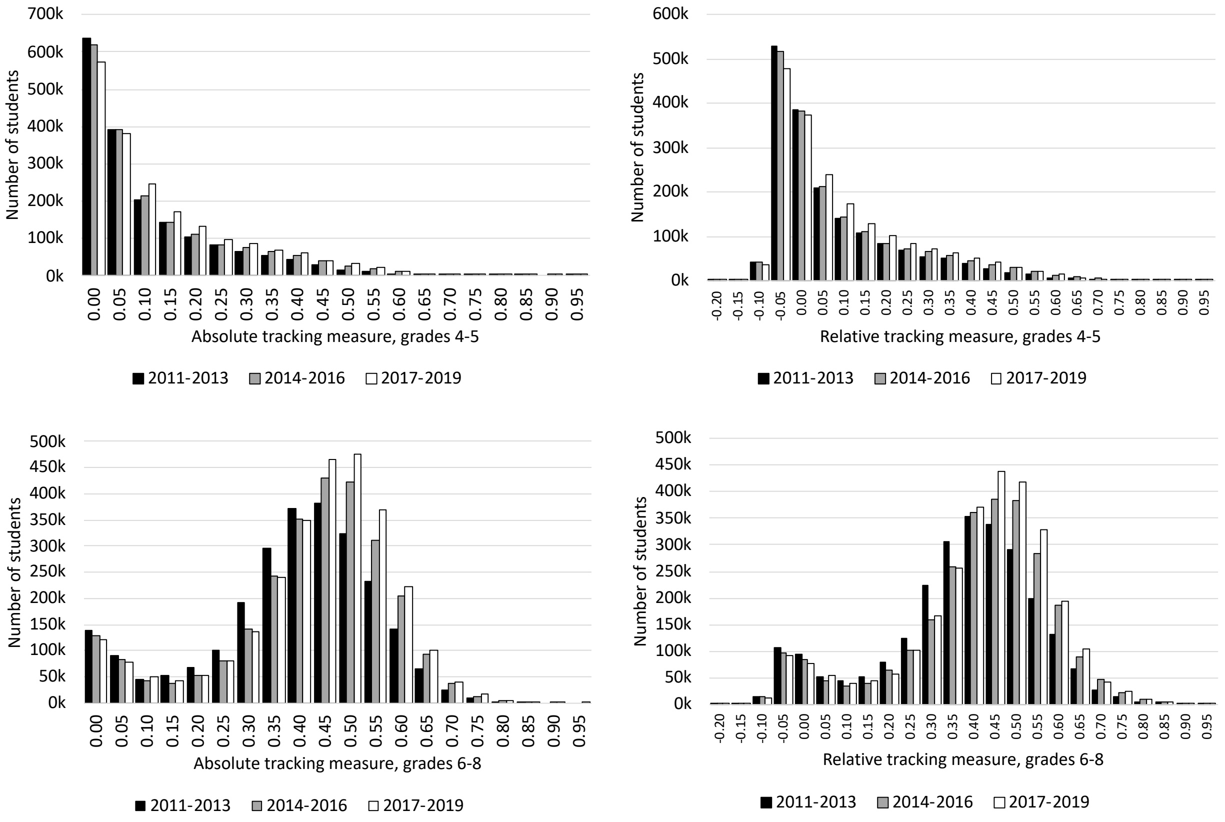
Appendix C. Supplementary Figures and Tables

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



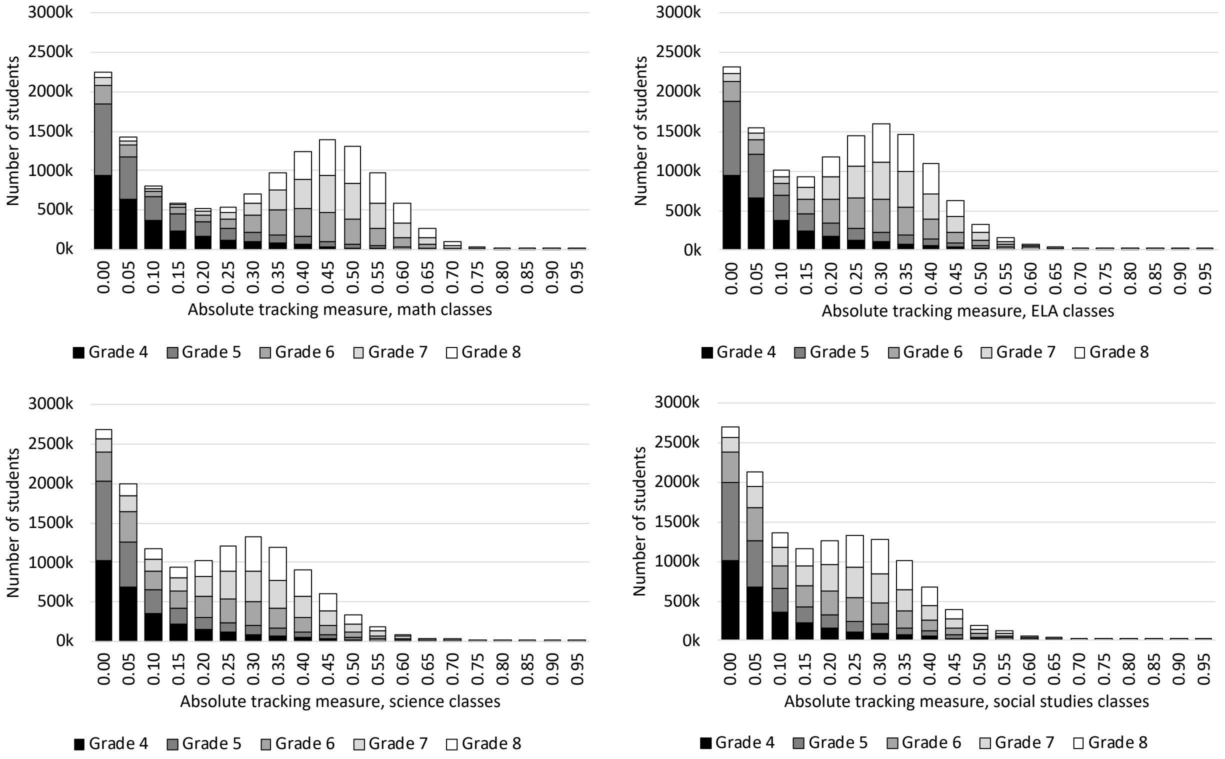
Notes: This figure shows the student-weighted distribution of the absolute tracking measure for students in middle school grades (6-8), broken down by whether the school serves any grades below grade 6.

Figure C2. Tracking over Time



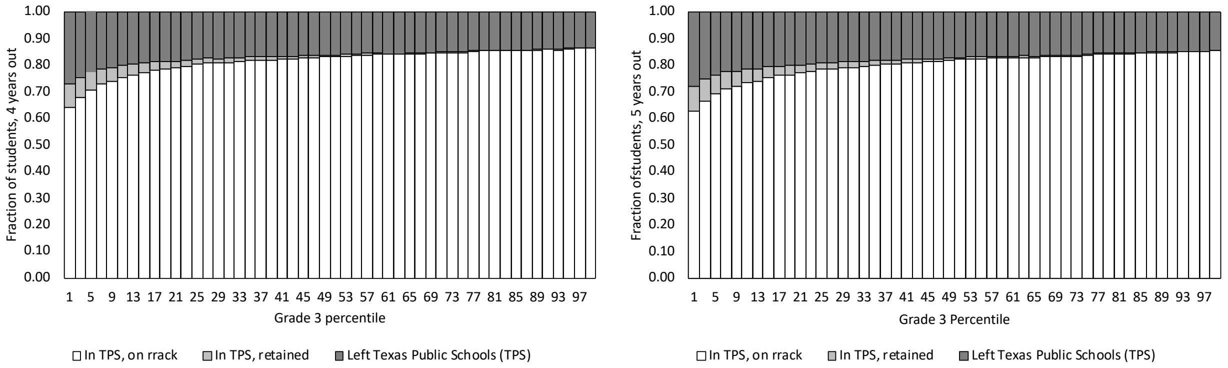
Notes: This figure shows the student-weighted distributions of the absolute and relative tracking measures, broken down by grade-level and time periods.

Figure C3. Absolute Tracking Measures for Math and Other Subjects



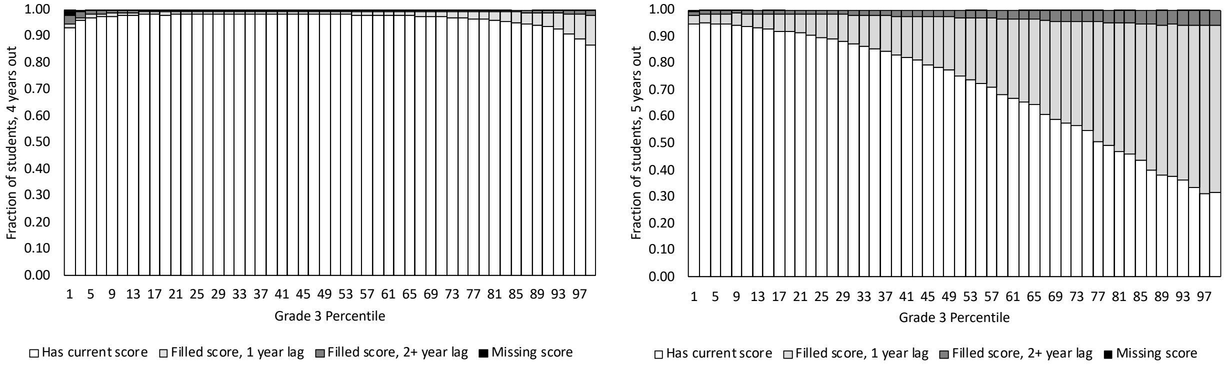
Notes: These panels show the student-weighted distribution of absolute tracking by prior math scores for math (top left), English language arts/reading (top right), science (bottom left), and social studies (bottom right) classes, broken down by grade.

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.