Ability Tracking, School and Parental Effort, and Student Achievement: A Structural Model and Estimation*

Chao Fu † Nirav Mehta ‡

March 23, 2017

Abstract

We develop and estimate an equilibrium model of ability tracking in which schools decide how to allocate students into ability tracks and choose track-specific teacher effort; parents choose effort in response. The model is estimated using data from the ECLS-K. Our model suggests that a counterfactual ban on tracking would benefit low-ability students but hurt high-ability students. Ignoring effort adjustments would significantly overstate the impacts. We then illustrate the tradeoffs involved when considering policies that affect schools' tracking decisions. Setting proficiency standards to maximize average achievement would lead schools to redistribute their inputs from low-ability students to high-ability students.

*We thank Yuseob Lee, George Orlov, and Atsuko Tanaka for excellent research assistance. We thank John Bound, Betsy Caucutt, Tim Conley, Steven Durlauf, John Kennan, Lance Lochner, Michael Lovenheim, Jesse Rothstein, Jeffrey Smith, Steven Stern, Todd Stinebrickner, Chris Taber, and Kenneth Wolpin for insightful discussions. We also thank the Editor and two anonymous referees for their comments. We have benefited from the comments from participants at the AEA 2014 winter meeting, Labo(u)r Day, and the CESifo Area Conference on the Economics of Education, and seminars at Amherst, ASU, Berkeley, Brock, Queen's, Rochester, St. Louis, and UWO.

†Department of Economics, University of Wisconsin. Email: cfu@ssc.wisc.edu.

‡Department of Economics, University of Western Ontario. Email: nirav.mehta@uwo.ca

1 Introduction

Ability tracking, the practice of allocating students into different classrooms based on prior performance, is pervasive. It is also controversial, because it may benefit students of certain ability levels while hurting others. There is considerable policy interest in learning how ability tracking affects different types of students and how policy changes, such as changing proficiency standards, would affect schools' tracking choices and the distribution of student achievement. However, the determinants and effects of tracking remain largely open questions.

Carefully-designed experiments can potentially answer these questions. However, to learn the effects of tracking, experiments would have to be conducted on a wide selection of schools, because, as Gamoran and Hallinan (1995) point out, the effectiveness of tracking depends on how it is implemented and on the school climate in which it is implemented. Similarly, because different policies may have very different impacts, one would need to run experiments under an extensive range of potential policies. The cost of doing so can become formidable very fast.

As a feasible alternative, we adopt a structural approach. Specifically, we develop and estimate a model that treats a school's tracking regime and track-specific effort, parental effort, and student achievement as joint equilibrium outcomes. The model is designed to address three interrelated components that have yet to be considered in a single framework. First, changing the peer composition of one classroom requires re-allocating students, necessarily changing peers in some other classroom(s). Therefore, it is important not to treat classrooms in isolation when studying the treatment effect of changing peers. Second, by explicitly modeling school and parental effort inputs, we can infer what these input levels and student achievement would be if tracking regimes, hence peer compositions, were changed. This allows us to decompose the effects of tracking into the direct effect of changes in peers and the indirect effects caused by the behavioral responses of schools and parents. As argued by Becker and Tomes (1976), ignoring behavioral responses may bias estimated impacts of policy changes. Finally, our explicit modeling of tracking regime choices allows us to predict how tracking regimes, which determine classroom-level peer composition, and subsequent school and parent inputs would change in response to a policy change.

In the model, each school is attended by children from different types of house-

holds, the distribution of which may differ between schools. A household type is defined by the child's ability and by how costly it is for the parent to help her child learn. A child's achievement depends on her own ability, effort inputs invested by the school and by her parent, and the quality of her peers.[1] A parent optimizes her child's achievement by choosing costly parental effort given both her child's track assignment, which determines peer quality, and the track-specific teacher effort.[2] A school's objective increases in the average achievement of its students and the fraction of students satisfying a proficiency standard. Taking into account responses by parents, the school chooses both a tracking regime and track-specific teacher effort inputs to maximize its own objective. In our model, a tracking regime is not a dichotomous choice; rather, it is a choice from many potential allocations of students into different tracks, where a track consists of students from a particular section of the school's student ability distribution. Therefore, the difference across track-specific student ability distributions measures the degree to which students are separated based on ability.

Our framework naturally allows policies to produce winners and losers because changes in tracking regimes at a school may differentially affect students of different ability levels and parental backgrounds. Moreover, because the model allows schools to base tracking decisions on the composition of households it serves, policies that affect schools' tracking decisions may create heterogeneous effects at the school level. That is, there will be a distribution of treatment effects within each school, and this distribution may differ across schools.

We estimate the model using data from the nationwide Early Childhood Longitudinal Study (ECLS-K), which are rich enough to allow us to model the interactions between schools and parents. Students are linked to their parents and teachers. For students, we observe prior test scores, class membership, and end-of-the-year test scores. Parents report the frequency with which they help their children with homework (parental effort). Teachers report the class-specific workload (school effort) and the overall level of ability among students in each of their classes, relative to other students in the same school. This rare last piece of information, which is available for

[1] See Epple and Romano (2011) for a recent review of the literature on peer effects in education.

[2] The importance of these types of input adjustments has been supported by evidence found in the literature. For example, Pop-Eleches and Urquiola (2013) and Das et al. (2013) both document changes in home inputs in response to quasi-experimental variation in school inputs.

multiple classes within a school, essentially allows us to "observe", or compute, two of the key components of the model. The first is the tracking regime used by each school, which we detect based on teacher reports. The second is the composition of peer ability in each track. Each track comprises a section of the school-specific ability distribution; the section is specified by the tracking regime. Assuming teacher reports accurately reveal tracking regimes, such information directly reveals peer quality, despite student ability not being directly observable and the endogenous grouping of students into tracks, as discussed in Betts (2011). Therefore, we are able to separate the roles of own ability and peer quality in the technology, a common identification concern in this literature.

We use the estimated model to conduct two policy evaluations. First, we quantify the effects of allowing ability tracking on the distribution of student achievement by comparing outcomes from the baseline model, where schools choose tracking regimes, with counterfactual outcomes where no schools are allowed to track students, i.e., all classes in a school have the same composition. Based on our tracking measure, over 95% of schools in our sample practice ability tracking under the baseline, which may be a higher fraction than reported in studies using a dichotomous tracking measure because we allow for small differences in track-level abilities in schools that track. These schools are affected by this policy differently depending on their existing degrees of tracking, which can be large or small. Overall, a tracking ban increases peer quality for low-ability students and decreases peer quality for high-ability students. Our results suggest that, in response, schools increase teacher effort on average, but parents react differently depending on their child's ability. The increase in the peer quality for low-ability students is effectively an increase in these households' endowments, causing their parents to reduce their provision of costly effort. Conversely, parents of high-ability students increase their effort because their children now have lower-quality peers. Altogether, our results suggest that banning tracking increases the achievement of students with below-median prior scores by 2.2% of a standard deviation (sd) in outcome test score and reduces the achievement of students with above-median prior scores by 4.2% sd. We find the average treatment effect from banning tracking is small, meaning the effects of tracking are mostly distributional in nature.[3]

[3]Bond and Lang (2013) note that plausible monotonic transformations of test scores can mitigate or even reverse estimated differences in gains for subgroups. Thus, we emphasize the distri-

It is worth noting, however, that peer quality is estimated to be a significant determinant of achievement: holding all other inputs constant and increasing peer quality by one standard deviation (sd) causes a 20% sd increase in the outcome test score, on average. The composite effect of increasing peer quality, which allows for behavioral responses by schools and parents, is less than half the size. For example, if the effort adjustments by schools and parents in response to the ban on tracking were ignored, one would overstate the loss for students with above-median prior scores by 121% and overstate the gain for students with below-median prior scores by 147%. The equilibrium nature of our framework helps to cast light on the open question of why estimated peer effects are typically small, as reviewed by Sacerdote (2011). It is exactly the fact that peer quality is an important input that induces substantial behavioral responses by schools and parents, which in turn mitigate the direct effect of peer quality.

Our second counterfactual experiment highlights the trade-offs involved in achieving academic policy goals, because policies that affect schools' tracking decisions will necessarily lead to increases in peer quality for some students and decreases for some other students. In the model, schools value both average achievement and the fraction of students testing above a proficiency standard, which policymakers may be able to control. In this counterfactual, we search for region-specific proficiency standards that would maximize average student achievement in each Census region. Achievement-maximizing standards would be higher than their baseline levels in every region, but do not increase without bound because the density of student ability eventually decreases, reducing the gain in average achievement due to increasing standards. Under these higher proficiency standards, schools would adjust their effort inputs and tracking regimes such that resources are moved from low-ability students to high-ability ones, leading to decreases in achievement for low-ability students and increases in achievement for high-ability students.

In interpreting our findings, we would like to stress that, as in other identification strategies in this literature, assumptions have to be made about the validity of our peer quality measure. Our approach implicitly assumes that teacher reports correctly measure the relative ranking of class-level student ability within the same school. Although this assumption is not directly testable, as student ability is only

butional effects of ability tracking rather than cross-group comparisons.

noisily measured by prior test scores, we can use these prior test scores to gauge the degree to which this assumption might be problematic. In particular, we test for differences in mean prior test scores for each pair of adjacent tracks and find that among the schools that track according to our measure 57% show no significant between-adjacent-track differences in prior scores. As such, our measure may pick up tracking that is too subtle to be viewed as tracking according to the traditional dichotomous definition. There are also 15 (out of 342) adjacent-track comparisons where teachers' reported rankings are significantly inconsistent with mean prior test scores. We conduct a set of robustness checks to alleviate concerns that potential track mismeasurement may play a significant role in our results.

Section 2 reviews related literature. Section 3 describes the model. Section 4 describes the data. Section 5 explains our estimation and identification strategy. Section 6 presents the estimation results and Section 7 presents results from our counterfactual exercises. Section 8 concludes. The appendix contains additional details about the data and model.

2 Related Literature

This paper brings together two strands of studies on ability tracking: one that measures how tracking affects student achievement and another that studies the determinants of tracking decisions.[4] We take a first step towards creating a comprehensive framework to understand tracking, by building and estimating a model where tracking regimes, track-specific inputs, parental effort, and student achievement are joint equilibrium outcomes.

There is considerable heterogeneity in results from empirical work assessing the effect of ability tracking on both the level and distribution of achievement. Argys et al. (1996) find that tracking reduces the performance of low-ability students, while Betts and Shkolnik (2000b) and Figlio and Page (2002) find no significant differences in student outcomes between tracked and untracked high schools, conditional on own ability. From an experiment in Kenya, Duflo et al. (2011) find that students of all abilities gain from tracking. Gamoran (1992) finds that the effects

[4]See Slavin (1987), Slavin (1990), Gamoran and Berends (1987), Hallinan (1990), Hallinan (1994) and Betts (2011) for reviews of studies in the first strand. Examples in the second strand include Gamoran (1992) and Hallinan (1992).

of tracking on high school students vary across schools. The heterogeneous findings from these studies indicate that there is no one "causal effect" of ability tracking that generalizes to all contexts, and therefore highlight the importance of explicitly taking into account that households differ within a school and that the distributions of households differ across schools, as do our model and empirical analyses.[5]

The primary focus of the tracking literature is on how changes in peer characteristics affect one's academic outcomes, which relates to the literature studying academic peer effects as reviewed in Sacerdote (2011). The majority of peer-effect studies estimate a reduced-form relationship between peer quality and own achievement, using, as in the tracking literature, either of two methods: 1) exploit random variation in peer group composition, which allows the researcher to study outcomes for the affected subset of students or 2) include fixed effects, the idea being that residual variation will be exogenous to own peer quality. Sacerdote (2011) notes that many studies find modest effects of peer quality using reduced-form linear-in-means models, but there is also a large degree of heterogeneity in their findings.[6] Sacerdote (2011) suggests that nonlinearities in the relationship between own and peer characteristics may help explain such heterogeneity, noting that several studies have found evidence of nonlinear peer effects in the reduced-form, wherein higher-ability students gain more from increases in peer quality than do lower-ability students (see, e.g., Hoxby and Weingarth (2005), Imberman et al. (2012), Burke and Sass (2013), and Lavy and Schlosser (2011).)

Our findings reinforce the importance of allowing for nonlinearity in such reduced-form regressions. We find that the structural test score production function is such that higher-ability students gain more from peer quality than do lower-ability students. However, we also show that restricting the production function to be linear would not significantly affect this paper's policy implications, because a nonlinear reduced-form relationship between peer quality and own outcomes would arise naturally from the nonlinear nature of school and parental responses to changes in peer

[5]There is also a literature studying between-school tracking, e.g., Hanushek and Woessmann (2006).

[6]Where possible, we have converted coefficient estimates from the literature to be in terms of changes in standard deviations of outcome scores, in terms of changes in standard deviation in peer quality (measured as the mean of their prior achievement). We were able to do so for Burke and Sass (2013), Hanushek et al. (2003), Kiss (2013), Lefgren (2004), and Vigdor and Nechyba (2007); the findings range from 0 to 0.10 sd increase in achievement resulting from a 1 sd increase in peer quality.

quality.

A closely related paper is Fruehwirth (2013), which identifies peer achievement spillovers using the introduction of a student accountability policy as an instrument.[7] She uses this instrumental variable to deal with the problem of nonrandom assignment, which is valid if students are not reassigned in response to the policy.[8] Our paper complements her work. Unlike Fruehwirth (2013), we assume that a student's achievement does not depend directly on the effort choices by peers, hence abstracting from direct social interactions within a class.[9] Instead, we model schools' decisions about student assignment and track-specific inputs to study the nonrandom assignment of students across classrooms and how parents respond.

Consistent with our findings, researchers have shown that parental investment responds to changes in other inputs. For example, Das et al. (2013) find evidence from India and Zambia that student test scores improve when schools receive unanticipated grants but not when grants are anticipated, i.e., households offset their own spending in response to anticipated grants. Using a regression discontinuity design to study the Romanian secondary school system, Pop-Eleches and Urquiola (2013) find that parents reduce their effort when their children attend a better school. Using data from the NELS, Houtenville and Conway (2008) find evidence suggesting decreases in parental effort in response to increased school resources. Liu et al. (2010) use the NLSY to analyze the interrelationships among school inputs, household migration, and maternal employment decisions. Their findings suggest that when parental responses are taken into account, changes in school quality may only have minor impacts on child test scores. Sacerdote (2011) notes that, given evidence supporting existence of peer effects, an important next step is to quantify the relative importance of peers, school, and home inputs. Our paper takes a first step towards filling this gap.

While our work focuses on how peer groups are determined within a school, a different literature studies how households sort into different schools. Epple et al. (2002) study how ability tracking by public schools may affect student sorting be-

[7] See Manski (1993), Moffitt (2001), and Brock and Durlauf (2001) for methodological contributions concerning the identification of peer effects and Betts and Shkolnik (2000a) for a review on empirical work in this respect.

[8] Assuming random assignment to classrooms within schools, Fruehwirth (2014) finds positive effects of peer parental education on student achievement.

[9] See Blume et al. (2011) for a comprehensive review of the social interactions literature.

tween private and public schools. They find that when public schools track by ability, they may attract higher ability students who otherwise would have attended private schools. This is consistent with our finding that high-ability students benefit from tracking. Caucutt (2002), Epple and Romano (1998), Ferreyra (2007), Mehta (2017), and Nechyba (2000) develop equilibrium models to study sorting between schools and its effects on peer composition. Our work complements this literature by studying schools' tracking decisions, which determine class-level peer groups faced by households within a school, and by emphasizing the interactions between a school and attendant households in the determination of student outcomes.

3 Model

In this section, we first introduce our theoretical framework, followed by more detailed model specifications; we defer discussions of our modeling choices until the end of the section. Each school is treated in isolation. A school makes decisions about ability tracking and track-specific inputs, knowing that parents will choose their own effort in response.

3.1 The Environment

A school s is endowed with a continuum of households of measure one. Households are of different types in that students have different ability levels (a) and parents have different parental effort costs ($z \in \{z_1, z_2\}$, where $z = z_1$ is the low-cost type). Student ability a is known to the household and the school, but z is a household's private information, which implies that students of the same ability are treated identically by a school. Let $g_s(a, z)$, $g_s(a)$ and $g_s(z|a)$ denote, respectively, the school-s specific joint distribution of household types, marginal distribution of ability, and conditional distribution of z given a. In the following, we suppress the school subscript (s) when it is not confusing to do so.

Throughout the paper, ability a refers to the characteristics of the student that affect her academic performance and is also the basis on which a student's school allocates her to a particular track. Researchers only observe a noisy measure of a.

3.1.1 Timing

The timing of the model is as follows:
Stage 1: The school chooses a tracking regime and track-specific effort inputs.
Stage 2: Observing the school's choices, parents choose their own parental effort.
Stage 3: Student test scores are realized.

3.1.2 Production Function

The achievement of student i in track j depends on the student's ability (a_i); peer quality, (q_j), which is the average ability of students in the same track; the coefficient of variation of ability of students in the same track $\left(q_j^{cv}\right)$; track-specific school effort $\left(e_j^s\right)$; parental effort (e_i^p); and a school-specific shifter (α_{0s}).[10] The test score y_{ji} measures student achievement with noise $\epsilon_{ji} \sim F_\epsilon$, which has density f_ϵ, such that

$$y_{ji} = Y(a_i, q_j, q_j^{cv}, e_j^s, e_i^p, \alpha_{0s}) + \epsilon_{ji}. \tag{1}$$

The school-specific shifter α_{0s} captures the idea that the production processes may differ across schools in ways that are not observed by the researcher. The vector $\{\alpha_{0s}\}_{s=1}^S$ is treated as a set of free parameters to be estimated, which is a very flexible way to introduce unobserved heterogeneity without imposing any distributional assumptions.[11] Such heterogeneity allows the model to capture *any* type of matching between households and schools, as the correlation between α_{0s} and the characteristics of households attending school s are entirely determined by the data.

3.2 Parent's Problem

A parent derives utility from her child's achievement and disutility from exerting parental effort. Given the track-specific school input $\left(e_j^s\right)$ and the peer quality (q_j) and coefficient of variation of peer ability $\left(q_j^{cv}\right)$ of her child's track, parent i chooses her own effort to maximize the utility from her child's achievement, net of her effort

[10] Arguably, the entire distribution of peer ability may matter. Following the literature, for feasibility reasons we have allowed only moments of the ability distribution, in our case the average and coefficient of variation of peer ability, to enter the production function.

[11] Though in principle identified, a more flexible technology, in which all production function parameters are allowed to differ between schools, would be beyond the limits of the data.

cost $C^p\left(e_i^p, z_i\right)$:

$$u(e_j^s, q_j, q_j^{cv}, a_i, z_i, \alpha_{0s}) = \max_{e_i^p \geq 0} \left\{ \ln\left(Y(a_i, q_j, q_j^{cv}, e_j^s, e_i^p, \alpha_{0s})\right) - C^p\left(e_i^p, z_i\right) \right\},$$

where $u(\cdot)$ denotes the parent's value function. Denote the optimal parental choice $e^{p*}(e_j^s, q_j, q_j^{cv}, a_i, z_i, \alpha_{0s})$. Notice that a parent's optimal choice depends not only on her child's ability a_i, her cost type z_i, and school intercept α_{0s}, but also on the other inputs in the test score production, including the ones chosen by the school (q_j, q_j^{cv}, and e_j^s).

3.3 School's Problem

A school cares about the average test score of its students and may also care about the fraction of students above a proficiency standard y^*. It chooses a tracking regime, which specifies how students are allocated across classrooms based on student ability, and track-specific inputs. Formally, a tracking regime is defined as follows.[12]

Definition 1 *Let $\mu_j(a) \in [0,1]$ denote the fraction of ability-a students assigned to track j, such that $\sum_j \mu_j(a) = 1$. A tracking regime is defined as $\mu = \{\mu_j(:)\}_j$.*

If no student of ability a is allocated to track j, then $\mu_j(a) = 0$. If track j does not exist, then $\mu_j(:) = 0$. Because all students with the same ability level are treated identically, $\mu_j(a)$ is also the probability that a student of ability a is allocated to track j. Note that the number of tracks can be computed from μ: $\sum_j \mathbf{1}\{\sum_a \mu_j(a) > 0\}$. The school's problem can be viewed in two steps: 1) choose a tracking regime; 2) choose track-specific inputs given the chosen regime. The problem can be solved backwards.

3.3.1 Optimal Track-Specific School Effort

Let $e^s \equiv \left\{e_j^s\right\}_j$ denote a school effort vector across the j tracks at a school. Given a tracking regime μ, the optimal choice of track-specific effort solves the following

[12] Another type of tracking, which we do not model, may happen within a class. We choose to focus on between-class tracking because it is prevalent in the data—according to our tracking measure, 95% of schools track students across classes. In contrast, when asked how often they split students within a class by ability levels, the majority of teachers report either never doing so or doing so less than once per week.

problem:

$$V_s(\mu) = \max_{e^s \geq 0} \left\{ \begin{array}{c} \int_i \left\{ \sum_j \left[E_{(z_i, \epsilon_{ji})} \left((y_{ji} + \omega \mathbf{1}\{y_{ji} > y^*\}) \, | a_i \right) - C^s(e_j^s) \right] \mu_j(a_i) \right\} di \\ s.t. \; y_{ji} = Y(a_i, q_j, q_j^{cv}, e_j^s, e_i^p, \alpha_{0s}) + \epsilon_{ji} \\ e_i^p = e^{p*}(e_j^s, q_j, q_j^{cv}, a_i, z_i, \alpha_{0s}) \\ n_j = \sum_a \mu_j(a) \, g_s(a) \\ q_j = \frac{1}{n_j} \sum_a \mu_j(a) \, g_s(a) \, a \\ q_j^{cv} = \frac{1}{q_j} \left(\frac{1}{n_j} \sum_a \mu_j(a) \, g_s(a) \, (a - q_j)^2 \right)^{1/2} \end{array} \right\}, \tag{2}$$

where V_s denotes the school's value function given regime μ. The parameter $\omega \geq 0$ measures the additional valuation a school may derive from a student passing the proficiency standard (y^*). $C^s(e_j^s)$ is the per-student effort cost in track j. The terms in the square brackets comprise student i's expected net contribution conditional on her being in track j (denominated in units of test scores), where the expectation is taken over both the test score shock ϵ_{ji} and the distribution of parent type z_i given student ability a_i, which is $g_s(z|a)$. In particular, a student contributes by her test score y_{ji} and an additional ω if y_{ji} is above y^*. Student i's total contribution to the school's objective is a weighted sum of her track-specific contributions, where the weights are given by her probabilities of being assigned to each track, $\{\mu_j(a_i)\}_j$; the overall objective of a school integrates over individual students' contributions. There are five constraints a school faces, listed in (2). The first two are the test score technology and the optimal response of the parent. The next three identity constraints show how the tracking regime μ defines the size (n_j), student quality (q_j), and coefficient of variation of ability (q_j^{cv}) of a track. Let $e^{s*}(\mu)$ be the optimal solution to (2).

3.3.2 Optimal Tracking Regime

A school's cost may vary with tracking regimes, captured by the function $D(\mu)$. Balancing benefits and costs of tracking, a school solves the following problem:[13]

$$\max_{\mu \in M_s} \left\{ V_s(\mu) - D(\mu) + \eta_\mu \right\},$$

[13]We assume the tracking decision is made by the school. In reality, it is possible that some parents may request that their child be placed in a certain track. We abstract from this in the model and instead focus on parental effort responses.

where η_μ is the school-specific idiosyncratic shifter associated with regime μ, which is i.i.d. across schools. M_s is the support of tracking regime for school s, which is specified in Section 3.5.2.

3.4 Equilibrium

Definition 2 *A subgame perfect Nash equilibrium in school s consists of* $\{e^{p*}(\cdot), e^{s*}(\cdot), \mu^*\}$, *such that*
1) For each $(e_j^s, q_j, q_j^{cv}, a_i, z_i, \alpha_{0s})$, $e^{p*}(\cdot)$ *solves the parent's problem;*
2) $(e^{s*}(\mu^*), \mu^*)$ *solves the school's problem.*[14]

We solve the model using backward induction. First, solve the parent's problem for any given $(e_j^s, q_j, q_j^{cv}, a_i, z_i, \alpha_{0s})$. Second, for a given μ, solve for the track-specific school inputs e^s. Finally, optimize over tracking regimes to obtain the optimal μ^* and the associated effort inputs $(e^{p*}(\cdot), e^{s*}(\cdot))$.[15]

3.5 Further Empirical Specifications

We present the final empirical specifications we have chosen after estimating a set of nested models, as will be further discussed at the end of this subsection.

3.5.1 Household Types

There are six types of households in a school, formed from two types of parents (low and high effort cost, respectively z_1 and z_2) and three school-specific student ability levels (a_1^s, a_2^s, a_3^s).[16] Household types are unobservable to the researcher but may be correlated with observed household characteristics x, which include a noisy measure of student ability (x^a) and parental characteristics, x^p, which are parental education and an indicator of single parenthood. Let $\Pr((a,z)|x,s)$ be the distribution of (a,z) conditional on x in school s. The joint distribution take the following form,

[14]The equilibrium at a particular school also depends on the distribution of households $g_s(a,z)$ and η_μ. We suppress this dependence for notational convenience.

[15]This involves evaluating a school's objective function at all possible tracking regimes.

[16]Specifically, we use a three-point discrete approximation of a normal distribution for each school, as described in Appendix A.1.

the details of which are in Appendix A.1:

$$\Pr\left(\left(a_l^s, z\right)|x, s\right) = \Pr\left(a = a_l^s | x^a, s\right) \Pr\left(z | x^p, a_l^s\right).$$

This specification has three features worth commenting on. First, ability distributions are school-specific and discrete. This assumption allows us to tractably model unobserved student heterogeneity in a manner that allows ability distributions to substantially vary between schools, which is important to our understanding why schools make different tracking decisions. Second, because pre-determined student ability is itself an outcome of earlier parental inputs, the latter being affected by parental types, the two unobserved characteristics of the households are inherently correlated with each other. We capture this correlation by allowing the distribution of parental types to depend not only on parental characteristics but also on student ability. Third, the noisy measure of student ability (prior test score) is assumed to be a sufficient statistic for ability, i.e., conditional on the prior test score, the ability distribution does not depend on parental characteristics. This means that parental characteristics may serve as shifters for parental effort, as they can be excluded from the production technology. We are not, however, assuming that parental characteristics are not informative about student ability, rather that the previous test scores summarize all the information. Indeed, pre-determined ability is itself likely affected by prior parental investments, as can be seen from the joint distribution of (x^a, x^p).

3.5.2 Tracking Regime

The support of tracking regimes (M_s) is finite and school-specific, and is subject to two constraints. First, the choice of tracking regimes in each school is constrained by the number of classrooms. Let K_s be the number of classrooms in school s; we assume that the size of a particular track can only take values from $\left\{0, \frac{1}{K_s}, \frac{2}{K_s}, \ldots, 1\right\}$. Schools with more classrooms have finer grids of M_s. Second, the ability distribution within a track cannot be "disjoint" in the sense that a track cannot mix low- and high-ability students while excluding middle-ability students. Subject to these two constraints, M_s contains all possible ways to allocate students across the K_s classrooms. If a track contains multiple classrooms $\left(n_j > \frac{1}{K_s}\right)$, the composition of

students is identical across classrooms in the same track.[17]

The cost of a tracking regime depends only on the number of tracks, and is given by

$$D(\mu) = \gamma_{|\mu|},$$

where $|\mu| \in \{1, 2, 3, 4\}$ is the number of tracks in regime μ. $\gamma = [\gamma_1, \gamma_2, \gamma_3, \gamma_4]$ is the vector of tracking costs, with γ_1 normalized to 0.

The permanent idiosyncratic shifter in the school objective function, η_μ, is unobserved to the researcher and follows an extreme-value distribution. From the researcher's point of view, conditional on a set of parameter values Θ, the probability of observing a particular tracking regime $\widetilde{\mu}$ in school s is given by

$$\frac{\exp\left(V_s(\widetilde{\mu}|\Theta)\right)}{\sum_{\mu'} \exp\left(V_s(\mu'|\Theta)\right)}. \tag{3}$$

3.5.3 Production Function

Student achievement is governed by

$$Y(a, q, q^{cv}, e^s, e^p, \alpha_{0s}) = \alpha_{0s} + a + \alpha_1 e^s + \alpha_2 e^p + \alpha_3 q + \alpha_4 e^s a + \alpha_5 e^s e^p + \alpha_6 \mathbf{1}\{a > q\} q + \alpha_7 e^s q^{cv}, \tag{4}$$

where the coefficient on one's own ability a is set to one.[18] The interaction terms α_4 and α_5 allow the marginal effect of school effort on student achievement to depend on student ability and parental effort, respectively. The interaction term α_6 allows the marginal product of peer quality to differ for students whose ability is higher than track mean ability. Finally, α_7 allows the coefficient of variation of peer ability q^{cv} to affect the marginal product of school effort. This allows teaching to be be more or less effective in classes with widely-heterogeneous students.[19]

[17] Notice that track sizes can be different, i.e., more students can be in one track than in another. However, we have abstracted from flexibility in the size of classes within tracks. There are mixed findings in the literature about the effect of class size on achievement, see, for example, Mishel et al. (2002) for a discussion.

[18] In other specifications where the coefficient on own ability is free, the estimate is not significantly different from one. Therefore, we choose to restrict it to be one in our final specification.

[19] We also estimated a specification where the coefficient of variation of peer ability also entered by itself, but dropped this term because it was estimated to be zero.

3.5.4 Cost Functions

The cost functions for both parent and school effort are assumed to be quadratic. The cost of parental effort is type-specific, where types differ in the linear coefficient $z \in \{z_1, z_2\}$ but are assumed to share the same quadratic coefficient (c^p), such that

$$C^P\left(e^p, z\right) = ze^p + c^p\left(e^p\right)^2.$$

The cost of school effort is given by

$$C^s\left(e^s\right) = c_1^s e^s + c_2^s (e^s)^2.$$

3.5.5 Measurement Errors

We assume that both the school effort e^s and the parent effort e^p are measured with idiosyncratic errors.[20] The observed school effort in track j $\left(\widetilde{e}_j^s\right)$ is given by

$$\widetilde{e}_j^s = e_j^{s*} + \zeta_j^s, \tag{5}$$

where $\zeta_j^s \sim N\left(0, \sigma_{\zeta^s}^2\right)$.
For parental effort, which is reported in discrete categories, we use an ordered probit model to map model effort into a probability distribution over observed effort, as specified in Appendix A.2.1. Let $\Pr(\widetilde{e}^p|e^{p*})$ denote the probability of observing $\widetilde{e}^p$ when model effort is e^{p*}.

3.5.6 Further Discussion of Model Specification

Nonlinear Peer Effects Although findings on peer effects have been mixed in the reduced-form literature, a recent common theme is that it is important to allow for a nonlinear relationship between peer quality and student achievement. As we show in our results, we do find the nonlinear terms in the production function (4) to be significant; this specification matches the data the best among various alternatives we have tried, as shown in Appendix D. However, using our estimated model with

[20] Todd and Wolpin (2003) discuss the implications of mismeasuring inputs when estimating production functions for cognitive achievement and Betts (2011) discusses this problem in the context of ability tracking. Stinebrickner and Stinebrickner (2004) show that it is important to account for measurement error in the study time of college students.

alternative specifications, including one with a linear production technology, we find robust counterfactual policy implications. This is because even if the production technology itself is linear, a reduced-form nonlinear relationship will arise naturally from the endogenous input responses, as illustrated in Appendix C.

More General Model Alternatives We have estimated a set of nested models including versions that are more general than the one presented to incorporate additional potentially important features.[21] We have chosen the relatively parsimonious specifications presented here because, in likelihood ratio tests, we cannot reject that the simpler model fits our data as well as the more complicated versions.[22] To be specific, on the household side, where we have allowed for additional heterogeneity in the effectiveness of parental effort, on top of the existing heterogeneity in parental effort costs. On the school side, we have allowed for 1) a nested logit counterpart of equation (3) in the school's regime choice, and 2) an additional payoff in the school's objective function that increases with the fraction of students achieving outstanding test scores.

4 Data

We use data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K). The ECLS-K is a national cohort-based study of children from kindergarten entry through middle school. Information was collected from children, parents, teachers, and schools in the fall and spring of children's kindergarten year (1998) and 1st grade, as well as the spring of 3rd, 5th, and 8th grade (2007). Schools were probabilistically sampled to be nationally representative. More than 20 students were targeted at each school for the kindergarten survey round. This results in a student panel which also serves as a repeated cross section for each school. The ECLS-K assessed student skills that are typically taught and developmentally

[21] The final choice of the detailed specification of the model is, of course, data-dependent. The more general models we have considered do not significantly complicate estimation or simulation. We view this as good news for future work using different data because one could feasibly extend our current specification.

[22] The conclusions from our counterfactual experiments are robust to the inclusion of these additional features (results available upon request).

important, such as math and reading skills. We focus on 5th grade reading classes.[23]

The data are rich enough to allow us to model the interactions between schools and parents. For students, we observe their prior (3rd grade) test scores (which are related to their ability), class membership (to identify their ability track), and end-of-the-year test scores, where test scores are results from the ECLS-K assessment.[24] Students are linked to parents, for whom we have a measure of parental inputs to educational production (frequency with which parents help their child with homework), and parental characteristics such as education and single-parenthood (which may affect parental effort costs).[25] Assuming that homework loads on students increase teachers' effort cost, we use homework loads reported by the teacher to measure the school's effort invested in each class.[26]

For the tracking regime, we exploit survey data containing teachers' reports on the ability level of their classes, which are available for several classes in the same school.[27] The question for reading classes is: "What is the reading ability level of this child's reading class, relative to the children in your school at this child's grade?" A teacher chooses one of the following four answers: a) Primarily high ability, b) Primarily average ability, c) Primarily low ability and d) Widely mixed ability.[28] We use the number of distinct answers given by teachers in different classes as the number of tracks in a school. Classes with identical teacher answers to this question are viewed as in the same track. The size of each track is calculated as the

[23]We focus on reading instead of math because reading has a much larger sample size.

[24]Notice that the noisy measure of student ability in our model is a student's test score in Grade 3, which could be viewed an outcome arising from decisions made in previous periods. Given our assumptions that the Grade-3 test score is a sufficient statistic for student ability and that schools track Grade-5 students based only on ability, treating the Grade-3 score as pre-determined is consistent with our framework.

[25]Ferreyra and Liang (2012) use time spent doing homework as a measure of student effort. Due to data limitations and computational complexity, we use one measure of school effort and one of parental effort, which are both presumably the most direct effort inputs in the production of reading skills. Admittedly, effort inputs could be multidimensional, which could be investigated with an extension of the model estimated on richer data.

[26]The same measure has been used in the study of the relationship between child, parent, and school effort by De Fraja et al. (2010). Admittedly, the use of homework loads as school effort is largely due to data limitations, yet it is not an unreasonable measure. Homework loads increase teachers' effort, who have to create and/or grade homework problem sets. Moreover, a teacher may face complaints from students for assigning too much homework. See Appendix A.2 for the survey questions.

[27]The ECLS-K follows many students at the same school. As such, we have the above information for several classes at each school.

[28]This variable was also used by Figlio and Page (2002) in their study of tracking.

number of classes in that track divided by the total number of classes. Although the relative ability ranking is a priori obvious between answers a), c) and b) or d), the relative ranking between b) and d) is less so. In schools where both b) and d) exist, on average, the mean prior test score for students in b) is higher; therefore, we assume b) is the higher track. On average, tracks ranked higher by our measure have students with higher mean prior scores and, hence, higher mean abilities.[29] We later discuss the extent to which our track rankings are consistent with prior scores.

Finally, the data indicate the Census region in which each school is located. We set proficiency cutoff y^* per Census region to match the proficiency rate in the data with that in the Achievement Results for State Assessments data. This data contains state-specific proficiency rates, which we aggregate to Census region.[30]

There are 8,853 fifth grade students in 1,772 schools in the ECLS-K sample. We delete observations missing key information, such as prior test scores, parental characteristics and track identity, leaving 7,332 students in 1,551 schools. Then, we exclude schools in which fewer than four classes or ten students are observed. The final sample includes 2,789 students in 205 schools. The last sample selection criterion costs us a significant number of observations. Given our purpose of studying the equilibrium within each school, this costly cut is necessary to guarantee that we have a reasonably representative sample for each school. Despite this sample restriction, Table 14 shows that, among the observed 5th-graders, the summary statistics for the entire ECLS-K sample and our final sample are not very different. A separate issue is sample attrition over survey rounds. The survey started with a nationally-representative sample of 16,665 kindergartners, among whom 8,853 remained in the sample by Grade 5. Statistics for the first survey round are compared between the whole sample and the sample of stayers in Table 15, which shows that the sample attrition is largely random.[31]

Like most datasets, the ECLS-K has both strengths and weaknesses compared with other data. Administrative datasets typically contain test score information for

[29] This holds for the mean prior score by track and also when we standardize mean prior scores by track, by dividing by the mean prior score at the school. See Tables 10-11 in Appendix B.

[30] `https://inventory.data.gov`. See Table 9 in this paper for regional cutoffs. Details are available upon request.

[31] The only slight non-randomness is that students who are more likely to remain in the sample are those with both parents (75% in the whole sample vs. 79% in the final sample) and/or college-educated parents (49% vs. 51%). Although it is comforting to see that the attrition is largely random, the slight non-randomness leads us to make some caveats, which are in Appendix G.

all students in a school, which has spurred a large literature estimating reduced-form social interactions models wherein one's own outcome is affected by both own and peer characteristics, as well as peer outcomes.[32] However, such datasets typically lack detailed measures of school and parental inputs, due to the fact that they are not based on surveys. It is precisely these input data that we use to study the endogenous interactions between the school and the parents. The tradeoff from using survey data is that they do not contain a large number of students per school. However, as far as we know, no currently available dataset, other than the ECLS-K, contains the information that is essential for estimating a model like ours. Moreover, Tables 14 and 15 suggest that our final estimation sample is largely representative of the ECLS-K sample, which itself is nationally representative.

4.1 Descriptive Statistics

The first row of Table 1 shows that over 95% of schools practice ability tracking, i.e., they group students into more than one track, according to our tracking measure. The most common number of tracks is three, which accounts for 46% of all schools. About 13% of schools have four tracks. To summarize the distribution of students across schools, for each school we calculate the mean and the coefficient of variation (CV) of student prior test scores and the fraction of students in the school whose prior scores were below the sample median. Rows 2-4 of Table 1 present the mean of these summary statistics across schools, by the number of tracks in the school. On average, schools with more tracks have higher dispersion and lower prior scores. For example, the average prior test score in schools with only one track is 53.4 with a CV of 0.139 and fewer than 45% of students below the median. In contrast, the average prior test score in schools with four tracks is 50.0 with a CV of 0.173 and more than 57% of students below the median.

Tables 2 and 3 present summary statistics by the number of tracks in the school and the identity of a track. For example, entries in Columns 7-8, Row 3 of Table 2 refer to students who belong to the third track in a school with four tracks. Table 2 shows that students in higher ability tracks have both higher average outcome test scores and a higher probability of passing the regional proficiency cutoff. Comparing average student outcome scores and pass rates in schools with one track to those

[32]This literature is discussed by Manski (1993), Moffitt (2001), and Sacerdote (2011).

Table 1: Student Prior Test Scores in Schools by Numbers of Tracks

	1 Track	2 Tracks	3 Tracks	4 Tracks	All Schools
% of schools	4.39	37.1	45.8	12.7	100.0
Mean	53.4	51.6	51.6	50.0	51.5
CV	0.139	0.165	0.160	0.173	0.163
% below median	44.1	48.3	49.8	57.1	50.0

in other schools, we see they are similar to those of the middle-track students in schools with three or four tracks. That is, the average student allocated to the lower (higher) track in a multiple-track school has poorer (better) outcomes than an average student attending a single-track school.

Remark 1 *One needs to look beyond the existence/non-existence of tracking or the number of tracks and look into the characteristics of each track in order to see the extent to which students are tracked. Evidence presented in Appendix E shows that, while the mean range in mean prior test scores (i.e., mean prior test score in a school's highest track minus that of the school's lowest track) is one standard deviation, this range varies among schools that track.*

Table 2: Average outcome score and percent of students passing the cutoff

	1 track		2 tracks		3 tracks		4 tracks	
Track	Score	% pass	Score	% pass	Score	% pass	Score	% pass
1	51.84	69.42	45.95	50.88	44.92	42.85	45.40	33.38
2			51.98	75.75	51.38	68.47	51.44	61.00
3					55.62	84.39	51.45	64.54
4							57.99	97.87
All	51.84	69.42	49.73	66.83	50.81	66.88	52.08	65.70

The top panel of Table 3 shows average school effort by track. With the exception of Track 4 in schools with four tracks, the average school effort (expected hours of homework done by students per week) increases with track level in schools with more than one track. At the school level, the average effort level stays roughly constant with the number of tracks. The bottom panel of Table 3 shows that parental effort shows the opposite pattern compared to school effort. Average parental effort

(frequency of helping child with homework in reading) decreases with student track level. For example, in schools with three tracks, while parents of lowest-track students on average help their children 2.57 times per week, parents of highest-track students do so only 2.11 times.

Table 4 summarizes parental effort and student outcomes by household characteristics. Compared to their counterpart, parents without college education, single parents and parents with lower-prior-achievement children exert more parental effort. Nevertheless, average student outcome test scores are lower in these households.

Table 3: Average teacher and parent effort by track and number of tracks

Teacher effort

Track	1 track	2 tracks	3 tracks	4 tracks
1	1.86	1.75	1.75	1.82
2		1.90	1.88	1.84
3			1.96	1.93
4				1.68
All	1.86	1.83	1.86	1.82

Parent effort

Track	1 track	2 tracks	3 tracks	4 tracks
1	2.07	2.31	2.57	2.29
2		2.03	2.37	2.71
3			2.11	2.78
4				2.08
All	2.07	2.11	2.27	2.38

Table 4: Parent effort and outcome test score by observed characteristics

	Parent effort	Outcome test score
Less than college	2.35	48.00
Parent college	2.12	54.24
Single-parent hh	2.37	48.76
Two-parent hh	2.18	52.37
Grade 3 score below median	2.61	45.35
Grade 3 score above median	1.82	57.96

4.2 Potential Misspecification of Tracking Regimes

We use teacher reports to detect tracking regimes. These reports permit identification of the role of peer quality, assuming they reveal the "true" tracking regime (see discussion in Section 5.2). The role they play in identification, combined with our focus on tracking regimes, means a discussion of the validity of the teacher report measure is necessary.

Although our tracking measure allows for variable track intensity, as we showed in the last section it also indicates that 95% of schools have more than one track, i.e., practice *some form of* ability tracking. Viewed through a dichotomous "tracking/no tracking" lens, this number may seem high. Therefore, we next examine the potential mismeasurement of tracking regimes. As is typically the case, student ability is not directly observed in our data, making it impossible to *directly* test the validity of our tracking measure. However, we can still assess whether our tracking regime measure is broadly consistent with the ranking of track-level student prior test scores, with the caveat that the latter are only noisy measures of student ability.

On average, tracks ranked higher by our measure have higher mean prior scores. However, this ranking could be violated within a school and, as was shown in Section 4.1, we also find evidence consistent with a large degree of heterogeneity in tracking intensity (Remark 1). This suggests it may also be informative to make within-school comparisons of mean prior scores by track, to further assess the validity of our tracking measure. Therefore, we test whether each track had a mean prior score at least as high as the track just below. Details of this analysis are in Appendix F.1.1; we summarize our findings here. At the 5% significance level, among the 342 pair-wise adjacent-track comparisons, 114 are significant with the higher track having a higher mean prior score, i.e., the "right" sign (Case 1), 139 are insignificant with the "right" sign (Case 2), 74 are insignificant with the higher track having a lower mean prior score, i.e., the "wrong" sign (Case 3) and 15 are significant with the "wrong" sign (Case 4).

Consistent with how tracking can be more or less intense in our model, we fail to reject there being significant differences in prior scores for students in different tracks in 57% schools we classify as practicing ability tracking. However, the existence, and number, of cases with the "wrong" sign indicate the presence of either potentially considerable measurement error of ability via prior scores and/or mis-

specification of tracking regimes. Therefore, we conduct extensive data analyses and robustness checks to examine this point. As shown in Appendix F, we find the following. First, statistics—in particular, correlations between endogenous variables—are similar across our original sample and subsamples excluding schools involving significant and/or insignificant "wrong"-sign cases. Second, we have re-estimated the model excluding i) any school with significant "wrong"-sign adjacent-track comparisons (i.e., Case 3) and ii) any school with (significant or insignificant) "wrong"-sign adjacent-track comparisons (Cases 3 and/or 4). The new estimates are similar to the original ones for all parameters in i) and most parameters in ii). Finally, we re-compute the ban-tracking counterfactual four times: a) for schools in Cases 1-3 only, b) for schools in Cases 1-2 only, both under the original estimates; c) and d) repeat the two exercises, but under the new estimates described in i) and ii), respectively. We find that the effects of banning tracking are essentially the same as the original ones under a)-c). Under d), the original results hold qualitatively, but the gain for below-median-prior-score students would be smaller and the loss for above-median-prior-score students would be larger.[33] These robustness checks mitigate our concerns about the potential misspecification of tracking regimes.

5 Estimation and Identification

5.1 Estimation

We estimate the model using maximum likelihood, where parameter estimates maximize the probability of observing the joint endogenous outcomes, given the observed distributions of household characteristics across schools.

The parameters Θ to be estimated include model parameters Θ^0 and parameters Θ^ζ that govern the distribution of effort measurement errors. The former (Θ^0) consists of the following seven groups: 1) Θ_y governing student achievement production function $Y(\cdot)$, which includes the vector of school-specific technology shifters $\{\alpha_{0s}\}_{s=1}^{S}$, 2) Θ_ϵ governing the distribution of shocks to test score ϵ, 3) Θ_{c^s} governing school effort cost, 4) Θ_{c^p} governing parental effort cost, 5) Θ_D governing the cost of tracking regimes, 6) ω, the weight associated with proficiency in school's objective

[33] The effects under d), versus the original effects, are 0.017 sd, vs. 0.022 sd for below-median-prior-score students, and -0.071 sd, vs. -0.042 sd for above-median-prior-score students.

function, 7) Θ_T governing the distribution $\Pr((a,z)|x,s)$ of household type given observables.[34]

The endogenous outcomes observed for school s (O_s) include the tracking regime $\widetilde{\mu}_s$, track-specific school effort $\left\{\widetilde{e}^s_{sj}\right\}_j$ and household-level outcomes: parental effort $\widetilde{e}^p_{si}$, the track to which the student is assigned τ_{si}, and student final test score y_{si}. Let $X_s = \{x_{si}\}_i$ be the observed household characteristics in school s. The vector X_s enters the likelihood through its relationship with household types (a,z), which in turn affect all of O_s.

The Likelihood The likelihood for school s is

$$L_s(\Theta) = l_{\widetilde{\mu}_s}\left(\Theta^0\right)\prod_j l_{sj}\left(\Theta\backslash\Theta_D\right)\prod_i l_{si}\left(\Theta_y,\Theta_\epsilon,\Theta_T,\Theta_{c^p},\Theta^\zeta\right),$$

where each part of the likelihood is as follows:

A. $l_{\widetilde{\mu}_s}(\Theta^0)$ is the probability of observing the tracking regime, which depends on all model parameters Θ^0, since every part of Θ^0 affects a school's tracking decision, but not on Θ^ζ. It is given by (3).

B. $l_{sj}(\Theta\backslash\Theta_D)$ is the contribution of the observed school effort $\widetilde{e}^s_{sj}$ in track j given the tracking regime $\widetilde{\mu}_s$. It depends on all Θ^0 but Θ_D since the latter does not affect school effort decision given the tracking regime. It also depends on Θ^ζ because school effort is measured with error:

$$l_{sj}(\Theta\backslash\Theta_D) = \frac{1}{\sigma_{\zeta^s}}\phi\left(\frac{\widetilde{e}^s_{sj} - e^{s*}_j(\widetilde{\mu}_s|X_s,\Theta^0\backslash\Theta_D)}{\sigma_{\zeta^s}}\right).$$

where ϕ denotes the standard normal density.

C. $l_{si}\left(\Theta_y,\Theta_\epsilon,\Theta_T,\Theta_{c^p},\Theta^\zeta\right)$ is the contribution of household i, which involves integrating type-specific contributions to the likelihood over the distribution of house-

[34]The distribution that enters the model directly, i.e., $g_s(a,z)$, does not involve additional parameters, because

$$g_s(a,z) = \int \Pr((a,z)|x,s)\,dF_s(x),$$

where $F_s(x)$ is the distribution of x in school s.

hold types:

$$l_{si}\left(\Theta_y,\Theta_\epsilon,\Theta_T,\Theta_{c^p},\Theta^\zeta\right)=\sum_{a,z}\Pr((a,z)\,|x_i,s,\Theta_T)l_{si}\left((a,z)\,|\Theta_y,\Theta_\epsilon,\Theta_{c^p},\Theta^\zeta\right),$$

where $l_{si}\left((a,z)\,|\Theta_y,\Theta_\epsilon,\Theta_{c^p},\Theta^\zeta\right)$ is the contribution of household i if it were type (a,z) :

$$l_{si}\left((a,z)\,|\Theta_y,\Theta_\epsilon,\Theta_{c^p},\Theta^\zeta\right)=\left[\begin{array}{l}\Pr\{track=\widetilde{\tau}_{si}|a,\widetilde{\mu}_s\}\times\\ \Pr(\widetilde{e}^p_{si}|e^{p*}(e^s_{\tau_{si}},q_{\tau_{si}},a,z|\Theta_y,\Theta_{c^p},\Theta^\zeta))\times\\ f_\epsilon\left[\left(y_{si}-Y(a,q_{\tau_{si}},e^s_{\tau_{si}},e^p(\cdot)|\Theta_y)\right)|\Theta_\epsilon\right]\end{array}\right].$$

The three components of $l_{si}\left((a,z)\,|\cdot\right)$ are
1) the probability of being assigned to $\widetilde{\tau}_{si}$ *given* tracking regime $\widetilde{\mu}_s$ *and* ability a, which is implied by $\widetilde{\mu}_s$ and $g_s\left(a\right)$;
2) the contribution of the observed parental effort $\widetilde{e}^p_{si}$ *given* peer quality and the model predicted school effort $e^s_{\tau_{si}}$ in track τ_{si}, which depends on parental cost parameters, the achievement parameters and the measurement error parameters; and
3) the contribution of test score *given* all model predicted inputs, which depends on achievement parameters and the test score distribution parameter.

5.2 Identification

Manski (1993) distinguishes between three components in social interactions models: endogenous effects (the direct effect of peer outcomes on one's own outcome), exogenous contextual effects (the effect of exogenous peer characteristics on one's own outcome), and correlated effects (shocks that are common to both one's peers and oneself). Separating these channels has been the focus of much empirical work.[35] The focus of our paper is to understand the interrelated nature of school tracking decisions, track input choices, and parental effort choices, not to push the frontier of identification of social interactions models. However, it does take these inferential problems seriously and addresses them by utilizing two arguments. First, similar to other work in this area (Caucutt (2002), Epple and Romano (1998), Epple et al. (2002)), the mechanism through which tracking operates is through peer quality.

[35]For example, Bramoullé et al. (2009) circumvents the reflection problem by exploiting sample finiteness. See Betts and Shkolnik (2000a) for a review of other work in this literature.

Although peer quality and one's own ability depend on the history of inputs, they are pre-determined at the beginning of the game. Therefore, as in this other work, the "reflection problem" of separating endogenous effects from exogenous contextual effects is not relevant here.[36] Although endogenous social interactions do not pose an identification problem in our context, we must address the problem of separating exogenous contextual effects, which operate through unobserved peer quality, and correlated effects. In our context, this is the same as accounting for selection into peer groups (Moffitt (2001)).

Intuitively, if we do not account for the correlation between own ability and peer quality when students are tracked, what looks like the effect of peers may instead be correlated unobservables (i.e., own ability and peer quality are correlated), creating an inferential problem. This concern may arise from two layers of unobservables: at the school level and within a school, at the track level. School-level unobservables can bias our findings if not controlled for because, for example, households with two working parents may buy into a better district with a more productive school but may have less time available for parenting. To address this concern, we introduce school-specific shifters α_{0s} in the production function, which also allow for any type of matching between households and schools. Given school-specific shifters, our identification of the production function then relies mainly on within-school variation.

To account for the correlation between own ability and peer quality within a school, we exploit two key pieces of information available in the ECLS-K survey. The first is students' prior test scores, which are assumed to be sufficient statistics for student abilities. The distribution of prior test scores in each school therefore maps into the school-specific ability distribution. The second piece of information comes from teacher surveys, which directly classify each student's track quality. This additional information, available for multiple classes, allows us to essentially "observe", or compute, two key components of model.[37] The first is the tracking regime used by each school. The second is the composition of peer ability in each track. Each track comprises a section of the school-specific ability distribution; the section is specified by the tracking regime. Such information directly reveals peer

[36] This also means we can estimate a linear-in-means technology without having to exploit non-linearities in peer effects, a strategy identified by Blume et al. (2006).

[37] Assuming we observe a representative sample of classes in each school.

quality, despite the fact that, as discussed in Betts (2011), a student's ability is not directly observable. Therefore, we are able to separate the roles of own ability and peer quality in the technology, a common identification concern in this literature.

5.3 Further Discussion of Ability

A few assumptions about student ability merit further discussion. First, it may seem to be a strong assumption that prior test scores are sufficient statistics for ability. For example, suppose a low-ability student, who happened to have a very high prior score, was observed in a low track and received a low outcome test score; if such an observation was not adjusted in some manner one could over-estimate the effect of tracking. However, note that the likelihood also considers the track assignment probability, $\Pr\{track = 1|a, \widetilde{\mu}_s\}$, which would be low for a high-ability student. Therefore, when the likelihood integrates over the ability distribution of this given student, it naturally downweights the unlikely case where the student was of high ability (as indicated by their high prior score) but assigned to a low track.

Second, we have followed the practice of most studies in this literature and treated test scores as (noisy) cardinal measures of ability (for prior test scores) or achievement (for outcome test scores). However, as noted by Bond and Lang (2013), test scores should most naturally be treated as ordinal. Caveats should be taken when one makes cross-group comparisons or average calculations of the effects on test scores, the magnitudes of which can be sensitive to the particular scale used in the test. Therefore, we view the distribution of outcomes reported in this paper, especially those in the counterfactual experiments, as more informative.

Third, we focus on one dimension of what may be multidimensional ability. Accordingly, our analysis focuses on one subject: reading. Admittedly, other dimensions of a student's ability, e.g., math and non-cognitive skills, may also affect her performance, and hence a school's tracking decision, in reading.[38] However, introducing multidimensional ability would substantially complicate the model and make computation intractable, because tracking in this case would involve dividing a multidimensional space (all students) into subspaces (tracks). We therefore follow most of the studies on peer effects, where one's achievement in a subject may be

[38]There is a growing literature on multidimensional human capital, which may encompass different dimensions of cognitive and non-cognitive skills, e.g., Cunha et al. (2010).

affected by peer ability in the same subject but not by that in other subjects.[39]

6 Results

6.1 Parameters

In this section, we present estimates of key parameters. Other parameter estimates can be found in Appendix B. The top panel of Table 5 presents the production technology parameter estimates.[40] We find that school effort and parental effort complement each other, while the interaction between a student's own ability and school effort is negative. To understand the magnitudes of these parameters, first we calculate outcome test scores according to the estimated production function by increasing one input at a time while keeping other inputs at their baseline equilibrium values—that is, not taking into account behavioral responses. The average marginal effect of increasing ability by one standard deviation (sd) above the model average causes a 72% sd increase in the outcome test score. On average, increasing peer quality by one sd causes a 20% sd increase in the outcome test score. We also find evidence of a heterogeneity in the marginal product of increasing peer quality: Increasing peer quality for a student whose ability is above the track average results in a 17% larger increase than doing so for a student with ability below the track average. On average, increasing school effort by one sd causes a 5% sd increase in the outcome test score; increasing parental effort by one sd increases the outcome test score by 61% sd. We do not estimate there to be a strong interaction between the track coefficient of variation in peer quality and the marginal product of school effort.

The ceteris paribus estimate of the effect of peer quality may seem larger than

[39] See Sacerdote (2011) for a survey of studies on peer effects. This assumption is also maintained in structural work, such as Epple et al. (2002)'s study of ability tracking.

This simplification is also supported by the data. In a regression of reading test score on the past reading scores of one's own and one's peers, parental and school effort inputs in reading, and parental characteristics (see the data section for the exact measures), the R squared is 0.71. The R squared increases marginally, to 0.72, when we add one's own past math score, the average past math score of one's peers, and parental and school effort inputs in math. This finding suggests that our results may not be significantly affected by our omission of math from the production of reading skills. However, it would still be an important extension to consider tracking in a multidimensional ability setting.

[40] Recall that the coefficient of student's own ability has been set to one.

some of the reduced-form estimates of peer effects found in the literature. However, reduced-form estimates correspond to the *composite effect* of peer quality, which is a combination of the direct effect of a change in peer quality and the indirect effect of input changes. Based on our estimates, this composite effect is such that 1 sd increase in peer quality would, on average, increase the outcome test score by 9% sd, which lies in the range of findings in the literature (see footnote 6 and Sacerdote (2011)). We discuss this result in more detail in Section 7.1.1.

Table 5: Key parameter estimates

Production technology parameters

Par.	Estimate	SE	
α_1	9.27	0.66	school effort
α_2	17.68	0.39	parent effort
α_3	0.28	0.04	track peer quality
α_4	-0.05	0.01	interaction: ability and school effort
α_5	1.71	0.30	interaction: school and parent effort
α_6	0.05	0.01	interaction: track peer quality and above track peer quality
α_7	-5.35	9.14	interaction: school effort and CV of track peer quality

School parameters

Par.	Estimate	SE	
ω	2.45	1.82	weight on passing proficiency standard
c_2^s	2.06	0.04	quadratic school effort cost
γ_2	-1.03	0.16	regime cost, 2 tracks
γ_3	-0.73	0.16	regime cost, 3 tracks
γ_4	0.41	0.23	regime cost, 4 tracks

Household cost parameters

Par.	Estimate	SE	
c^p	0.10	0.003	quadratic parent effort cost
z_1	0.08	0.01	linear parent effort cost, low cost type
z_2	0.11	0.01	linear parent effort cost, high cost type
θ_0^c	-44.02	2.19	cost type intercept
θ_1^c	0.85	0.02	cost type, ability
θ_2^c	-10.74	3.38	cost type, single parent indicator
θ_3^c	3.62	2.45	cost type, college indicator

Instead of presenting the estimates for all the school-specific shifters, we summarize the correlation between the shifters and attendant household characteristics via an OLS regression. The results, shown in Table 13 in Appendix B, show

that school-specific shifters are positively correlated with student prior test score, parental education and the presence of both parents, suggesting positive assortative matching.

The middle panel of Table 5 presents estimates of school-side parameters. We estimate a low value for ω, suggesting that schools do not care much about how many students are proficient, compared to their concerns about average achievement. This finding may be due to the fact that the test score we use is an ECLS-K survey instrument, not a high-stakes test. This is consistent with findings from the school accountability literature, which finds that pressure, such as No Child Left Behind (NCLB), leads to large gains on high-stakes tests, but much smaller gains on low-stakes exams.[41] The second row shows that the cost of school effort is convex in effort levels.[42] The last three rows show that the cost of tracking regimes is non-monotone in the number of tracks. The most costly tracking regimes are those with four tracks (the highest possible number), followed by those with only one track (the cost of which normalized to zero). Without further information, our model is unable to distinguish between various components of the cost associated with different tracking regimes. However, we think these estimates are not unreasonable. On the one hand, increasing the number of tracks may involve developing more types of curricula as well as incur higher resistance from parents. On the other hand, pooling all students into one track may make the classroom too heterogeneous and thus difficult for the teacher to handle or it may be difficult to ensure that there are no differences between the realized distribution of student ability between classrooms. If these competing costs are both convex in the number of tracks, one would expect the total cost to be higher for the one- and four-track regimes. governing parental effort cost and the probability of being a high-cost parent. We find that the cost of parental effort is convex. The linear type-specific cost term is 43% higher for the high-cost type than for the low-cost type. Evaluated at the baseline levels of parental effort, the high-cost type incurs a 13% higher cost than the low-cost type. The last four rows show the relationship between household characteristics with parental cost types. Parents with higher education and higher-ability children are more likely to have

[41] See, for example, Koretz and Barron (1998), Linn (2000), Klein et al. (2000), Carnoy and Loeb (2002), Hanushek and Raymond (2005), Jacob (2005), Wong et al. (2009), Dee and Jacob (2011), and Reback et al. (2014).

[42] We cannot reject that the linear cost term c_1^s is zero, so we set it to zero.

higher effort costs, as are two-parent households. This is consistent with the data in Table 4, where parents without a college education and parents of children with lower prior achievement exert more parental effort, yet average student outcome test scores are lower in these households.

6.2 Model Fit

Overall, the model fits the data well. Table 6 shows model fit in terms of tracking patterns. The first two columns show that the model closely matches the distribution of tracking regimes across all schools. The next two columns show the fit for schools with lower spread of prior test scores, where a school is called a low-spread school if it has a coefficient of variation (CV) of prior scores below the median CV. The model slightly overpredicts the fraction of two-track schools and underpredicts that of three-track schools. The next two columns focus on schools with higher-ability students. In particular, we rank schools by the fraction of lower-prior-score (below median score) students from high to low: the higher the ranking of a school, the higher fraction of its students are of low prior scores, meaning they are more likely to be of low ability. We report the tracking regime distribution among schools that are ranked below the median in this ranking, i.e., schools with relatively better students. Overall, the model captures the pattern that schools with less variation and/or higher prior test scores are more likely to have only one track and less likely to have four tracks. The top panel of Table 7 shows that the model fits

Table 6: Tracking regimes

	All schools		Low Spread*		Low fraction of low ability**	
	Data	Model	Data	Model	Data	Model
% 1 track	4.39	4.96	4.85	5.26	4.55	5.16
% 2 tracks	37.07	36.96	36.89	38.86	40.91	38.58
% 3 tracks	45.85	46.62	47.57	44.56	47.27	44.88
% 4 tracks	12.68	12.46	10.68	11.32	7.27	11.38

* "Low spread" schools have a below-median coefficient of variation in prior score.

** "Low fraction of low ability" schools have a below-median fraction of schools with below-median prior score.

average outcome scores by track. The exception is that the model over-predicts the average score in the second track within two-track schools and that in the third

track within four-track schools. Figure 1 shows the model predicted CDF of outcome test scores contrasted with the data counterpart. The left panel shows the case for all students. The right panel splits the distribution by parental education and by single-parenthood. As seen, the model-predicted score distributions match well with the ones in the data.

Table 7: Outcomes by track and number of tracks

Mean outcome test score

	1 Track		2 Tracks		3 Tracks		4 Tracks	
Track	Data	Model	Data	Model	Data	Model	Data	Model
1	51.84	50.97	45.95	47.70	44.92	46.10	45.40	46.48
2			51.98	54.95	51.38	51.22	51.44	50.55
3					55.62	56.27	51.45	54.65
4							57.99	58.14

Mean school effort

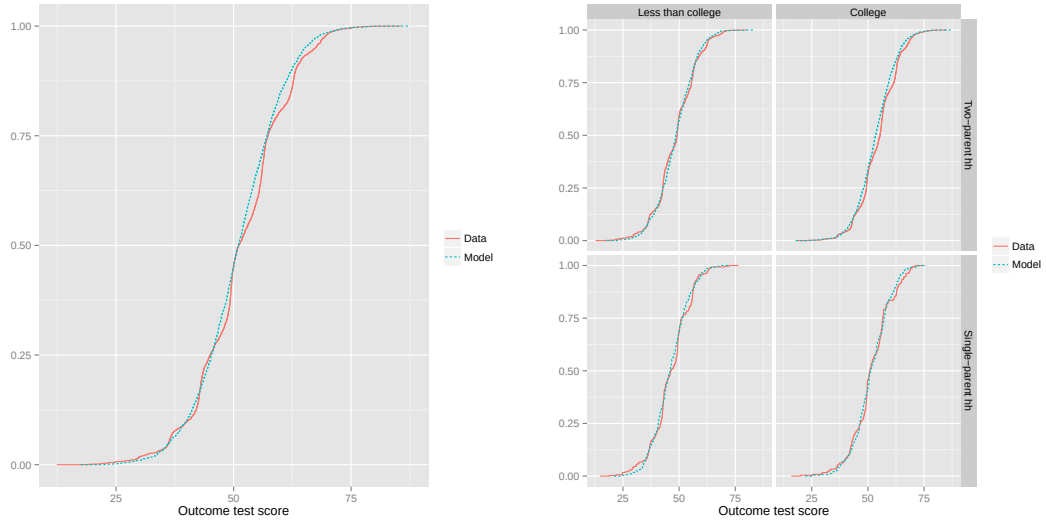
	1 Track		2 Tracks		3 Tracks		4 Tracks	
Track	Data	Model	Data	Model	Data	Model	Data	Model
1	1.86	1.87	1.75	1.88	1.75	1.87	1.82	1.87
2			1.90	1.83	1.88	1.87	1.84	1.89
3					1.96	1.81	1.93	1.85
4							1.68	1.81

Mean parent effort

	1 Track		2 Tracks		3 Tracks		4 Tracks	
Track	Data	Model	Data	Model	Data	Model	Data	Model
1	2.07	2.21	2.31	2.46	2.57	2.63	2.29	2.76
2			2.03	1.96	2.37	2.26	2.71	2.46
3					2.11	1.93	2.78	2.17
4							2.08	1.95

The middle panel of Table 7 shows the model fit for school effort. Compared to the data, the model predicts a flatter profile for school effort across tracks. The bottom panel of Table 7 shows the model fit for mean parental effort: the model matches the fact that parent effort decreases with track level, although the gradient is over-predicted in the four-track case (last two columns). Finally, Table 8 shows that the model fits well the level of parental effort and outcome scores by household characteristics. Parents with less education, single parents, and students with lower

Figure 1: Fit: Outcome test scores



prior score all have higher parent effort levels and lower outcome test scores.

Table 8: Means of parent effort and outcome score, by household characteristics

	Parent effort		Outcome score	
	Data	Model	Data	Model
Less than college	2.35	2.39	48.00	48.28
College	2.12	2.17	54.24	53.37
Single parent	2.37	2.38	48.76	48.89
Two parents	2.18	2.23	52.37	51.84
Low prior score	2.52	2.50	45.35	47.46
High prior score	1.78	2.02	57.96	55.09

7 Counterfactual Simulations

We use the estimated model to simulate two policy-relevant counterfactual scenarios. We contrast the outcomes between the baseline and each of the counterfactual cases. In particular, we present Average Treatment Effects (ATE) for different endogenous outcomes (some of which are inputs to achievement), for subgroups of students defined by their characteristics, such as prior test scores.[43]

[43]When a continuous variable is used to define subgroups, as they are in Figures 3 and 4(a), ATEs are calculated by non-parametric smoothing.

In the first counterfactual simulation, we quantify the effect of allowing tracking by solving the model when tracking is banned (i.e., all schools have only one track). We compare the changes in school effort, parental effort and student achievement. Our results indicate that failing to account for the equilibrium interactions between schools and parents could substantially bias the results. In the second counterfactual simulation, we examine the equilibrium effects of prospective changes in proficiency standards. In particular, we solve for optimal region-specific proficiency standards that would maximize average achievement. Unlike the tracking-ban counterfactual, here a school re-optimizes its tracking decision in response to this policy change.[44]

It is worth noting that the first counterfactual simulation enables one to understand the effects of any tracking, regardless of how intensely different schools choose to track when allowed to do so under the baseline. To implement such a policy, a technology that detects and monitors tracking practice would be required. Such a technology is not necessary for the second counterfactual policy.

7.1 Heterogeneous Effects of Tracking

We compare the outcomes under the baseline with those when tracking is banned and every school has to put all of its students into a single track. According to our tracking measure, over 95% of the simulated schools practice ability tracking to some degree under the baseline, hence are affected by this counterfactual and experience an exogenously imposed change in peer quality within classrooms.[45]

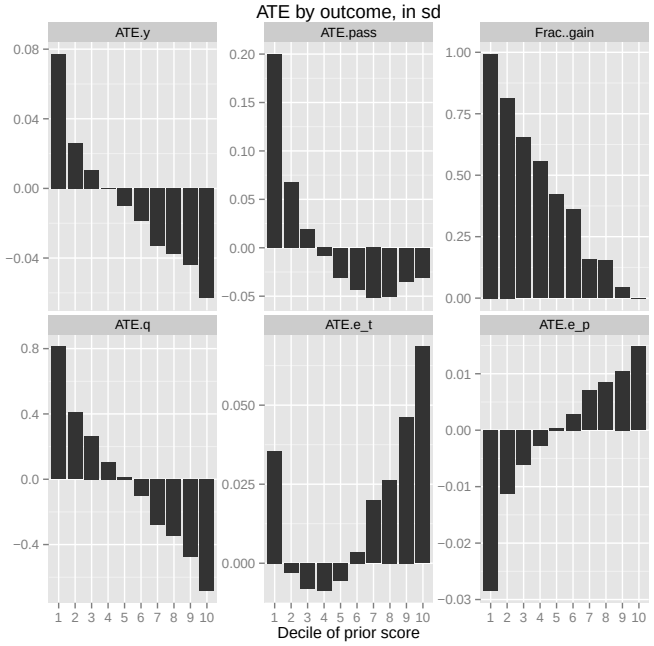
The first two panels in the top row of Figure 2 show average results for outcome test scores and pass rates, by decile of prior test score. The effects of a tracking ban are positive for students with lower prior test scores and negative for those with higher prior test scores, as measured in both the level of the final test scores and the pass rates. In particular, students with below-median prior scores gain 2.2% sd when ability tracking is banned, while those with above-median prior scores lose 4.2% sd when tracking is banned.[46] Consistent with the first two panels, the third

[44]Results from our counterfactual experiments are subject to the caveat that household distributions across schools are fixed. Incorporating households' school choices into our framework is an important extension we leave for the future work.

[45] Appendix F shows that these results are robust to the potential misspecification of tracking regimes.

[46]The ATE of banning tracking, over all students, is -0.8% sd in test scores. Our result that tracking (in the short run) hurts lower-ability students and benefits higher-ability students is

Figure 2: Change in outcomes and inputs due to banning tracking, by decile prior score



panel shows that the fraction of students who gain from a tracking ban declines with prior test scores.

Underlying the changes in student outcomes are the changes in the inputs, i.e., peer quality, school effort and parental effort, which are plotted in the bottom three panels of Figure 2, by decile of prior test scores. The tracking ban places all students in a school in one track, which means that lower ability students are on average placed with better students, and are made better off through the technology, ceteris paribus. The opposite holds for higher ability students. However, that is not the entire story. Both school effort and parental effort adjust to the change in peer composition imposed by this policy. Without the freedom to optimize over tracking regimes, schools that used to track students can only optimize over their effort inputs. As there is only one track, a school can only choose one effort level for all students. On average, schools increase their effort for students in most deciles (the second panel on the bottom). This change is most obvious for students with very high or very low prior test scores, who are more likely to have been tracked under

consistent with findings from some previous studies, e.g., Betts and Shkolnik (2000b) and Hoffer (1992).

the baseline. Unlike schools, which choose one effort level for all students in one track, parents can always adjust their effort levels for their own children. Indeed, the last panel shows that changes in parental effort are not only quantitatively but also qualitatively different across students with different prior test scores. Average parental effort decreases for students below the median and increases for those above the median. In particular, parents of students in the highest decile increase their inputs by the largest amount when tracking is banned, by about 1.5% sd.

Figure 2 highlights the trade-offs a school faces when choosing a tracking regime. Improving peer quality in one track necessarily involves reducing it in another, which will in turn lead to changes in parental effort. When low-ability students are grouped with higher-ability students, peer quality increases (decreases) for students with low (high) ability. For parents of low-ability students, who have been exerting much higher effort than those with high-ability students (Table 8), the exogenous increase in peer quality provides strong incentives for them to reduce their own effort. For parents with high-ability students, the exogenous decrease in peer quality pushes them to increase their own effort as a remedy. However, given the concavity of the parents' net payoff with respect to their own effort, the reduction in effort by the parents of low ability children is larger than the increase in effort by the parents of high ability children, especially among parents with very low-achieving children. To curb the reduction in parental effort among low-achieving households and encourage more effort in high-achieving households, the school increases its own effort, utilizing the fact that school effort and parental efforts are complementary to each other. Depending on the different sets of households they face, schools differ in how much effort they need to exert and how their students perform when they do not track versus when they track. These differences drive schools' different tracking decisions.

7.1.1 The Importance of Accounting for Behavioral Changes: Effort Inputs

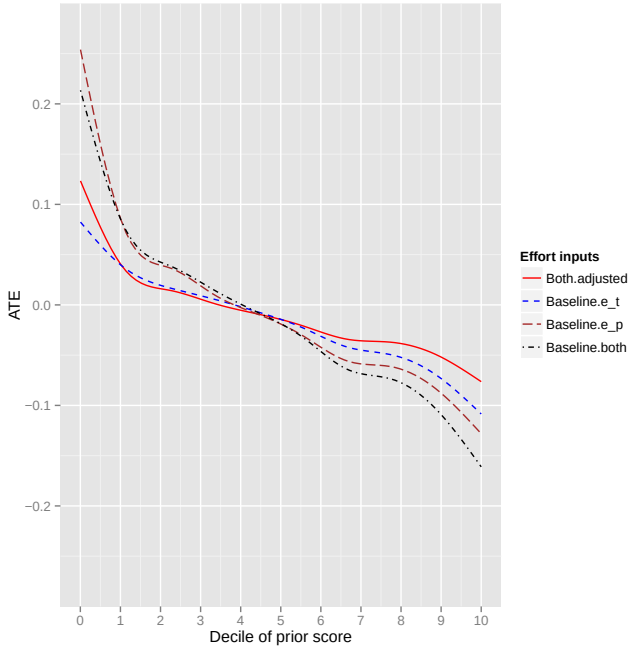
The test score technology plays an important role in evaluating the effect of tracking on student outcomes. This might prompt one to ask whether estimates of parameters governing the technology alone would adequately characterize outcomes for students were tracking banned. To illustrate the value of estimating an equilibrium model

where schools and parents may respond to changes in track peer quality, we contrast the ATE of banning tracking from our model prediction with the ATE ignoring endogenous effort responses. Let $(q^*, q^{cv*}, e^{s*}, e^{p*})$ denote inputs under the baseline scenario where tracking is endogenous, and $(q^*_{\text{CF}}, q^{cv*}_{\text{CF}}, e^{s*}_{\text{CF}}, e^{p*}_{\text{CF}})$ denote counterfactual equilibrium inputs when tracking is banned.

Figure 3 graphs the ATE for test scores (y-axis) against student prior test scores (x-axis). The red (solid) line is the model-predicted ATE, taking into account the change in peer quality induced by a tracking ban, as well as school and parent effort responses, i.e., $Y(a, q^*_{\text{CF}}, q^{cv*}_{\text{CF}}, e^{s*}_{\text{CF}}, e^{p*}_{\text{CF}}) - Y(a, q^*, q^{cv*}, e^{s*}, e^{p*})$ (recall $Y(\cdot)$ is the test score technology). The blue (short dashed) line is the ATE, ignoring school effort adjustments, i.e., $Y(a, q^*_{\text{CF}}, q^{cv*}_{\text{CF}}, e^{s*}, e^{p*}_{\text{CF}}) - Y(a, q^*, q^{cv*}, e^{s*}, e^{p*})$. Banning tracking pushes schools to increase their effort; ignoring this biases the effect of tracking ban downwards, although not by much. The bias from ignoring parental responses is much larger. The brown (long dashed) line is the ATE for test scores ignoring parental effort adjustment, i.e., $Y(a, q^*_{\text{CF}}, q^{cv*}_{\text{CF}}, e^{s*}_{\text{CF}}, e^{p*}) - Y(a, q^*, q^{cv*}, e^{s*}, e^{p*})$. When tracking is banned, lower-achieving students receive more inputs from the school, in terms of both peer quality and school effort. In response, parents of these students reduce their own effort. Failing to take into account this reduction drastically overstates the ATE of banning tracking for these students. The brown line lies far above the red line for students with the lower prior scores, especially those at the end of the distribution. The opposite is true for students with higher prior scores, whose parents increase their provision of costly effort in response to the lower peer quality. The black (dotted-dashed) line is the ATE for test scores ignoring both school and parent effort adjustments, i.e., $Y(a, q^*_{\text{CF}}, q^{cv*}_{\text{CF}}, e^{s*}, e^{p*}) - Y(a, q^*, q^{cv*}, e^{s*}, e^{p*})$. On average, ignoring effort changes would cause one to overstate the gains from banning tracking by 147% for students with below-median prior scores, and overstate the loss by 121% for students with above-median prior scores.

We have repeated the above counterfactual experiments using the estimated model with alternative production technology specifications, including one with a linear technology; the detailed results are shown in Appendix Table 16. Across all these specifications, the message remains the same: it is important to account for effort responses. For students with below-median prior scores, the bias from ignoring effort responses ranges from 70%-121%. For students with above-median prior scores, the bias from ignoring effort responses ranges from 102%-147%.

Figure 3: ATE of banning tracking : equilibrium vs. fixed effort inputs



Two points are worth noting. First, although the effect of a tracking ban on test scores is significantly attenuated by parental effort responses, its effect on household welfare is likely to be larger. In particular, banning tracking benefits households with lower-prior-score children both by increasing student achievement and by saving parental effort; while it hurts households with higher-prior-score children both by lowering student achievement and by increasing parental effort. Second, we study the effect of (non)tracking while holding the distribution of households across schools constant. It is plausible that some households will react by changing schools. For example, some households with high-ability children may change schools to reduce their losses from a ban on tracking, which will in turn decrease the gain for low-ability students from the ban. That is, taking this additional layer of household response into account may further attenuate the policy effect.

7.2 Optimal Proficiency Standards: Trade-offs

Policy changes will often create winners and losers. This is especially true in the case of educational policies that may affect ability tracking, which has qualitatively different effects on students of varying ability levels. This is because changing peer

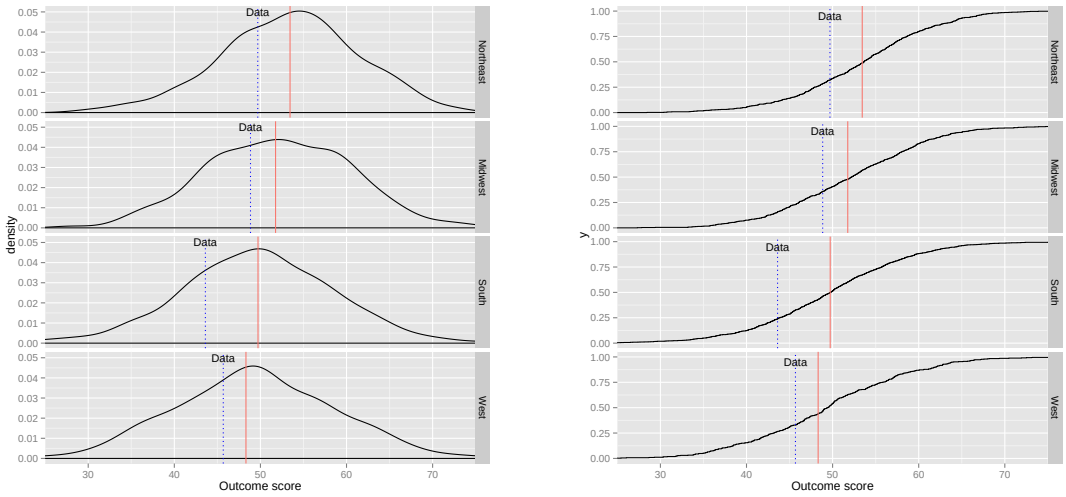
quality for some students necessarily involves changing peer quality for some other students, which is accompanied by adjustment in the effort choices by the school and by parents. By incorporating tracking regimes, school effort, and parental effort inputs into one framework, our work lends itself to a better understanding of the effects of education policy, especially with respect to the trade-offs they involve. Our second counterfactual experiment highlights this point by examining how changes in proficiency standards would affect the distribution of student achievement. Because schools care about the fraction of students above the proficiency standard, changes in these standards can shift outcomes toward certain policy goals.[47] To examine these trade-offs, we search for region-specific proficiency standards that maximize the region-specific averages of student achievement. Note that this is only one illustration of trade-offs; one could also use our framework to study the effects of other policy changes, which may generate different distributions of winners and losers.

Figure 4 places proficiency standards in relation to the distribution of baseline outcome test scores, by region, where the left panel (4(a)) shows the density and the right panel (4(b)) shows the CDF. Each panel overlays the distribution of baseline outcome scores with the baseline proficiency standards (blue dotted line). Loosely speaking, the ranking of the regional distributions of student baseline achievement from high to low (in the sense of first order stochastic dominance) is the Northeast, the Midwest, the West, and lastly the South. This ranking lines up with that of regional proficiency standards, with the Northeast having the highest standard and the South having the lowest. In all four regions, the baseline (data) standard is approximately located at the 30th percentile of the regional distribution of the outcome scores.

The regional standards that maximize region-specific average achievement are the red solid lines in Figure 4, which are higher in all four regions than the baseline standards. When standards are lower (as in the baseline), schools have the incentive to improve outcomes near the lower end of the distribution, which would sacrifice a much larger measure of students near the middle and top of the distribution of baseline scores. The same argument applies for the case if standards are too

[47] As mentioned earlier in the paper, our estimate of a school's preference for the lower-tail of the score distribution does not reflect the pressure it faces from high-stakes tests; thus, this experiment should not be interpreted as increasing the bar for a high-stakes test.

Figure 4: Proficiency standards by region, data and counterfactual



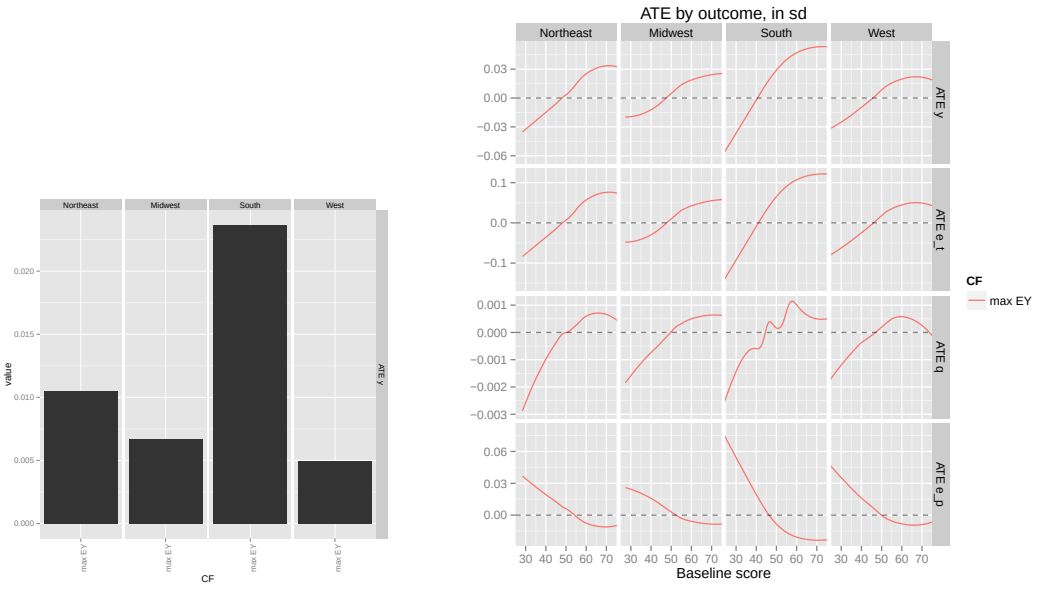
(a) Density of outcome scores

(b) CDF of outcome scores

high. The new standards maximize the average performance by moving schools' attention away from the low-performing students toward a location that rewards schools for improving mean test scores. In fact, the new standards are located at around the median of each region's baseline distribution of test scores, which is also where the density of outcome scores is highest, as seen in Figure 4(a). Figure 5(a) summarizes the ATE on outcome scores by region due to the change in standards. Setting standards to maximize average achievement has the biggest ATE in the South, followed by the Northeast, the Midwest and finally the West. Intuitively, the ATE is increasing in the difference between baseline and achievement-maximizing standards.

Figure 5(b) plots the ATE on the outcome score and inputs (school effort, peer quality, and parental effort) by baseline score and region. The top panel shows that in all four regions, the ATE on outcome test scores increases with student baseline outcome scores. In fact, the ATE is negative for students with low baseline scores and positive for students with high baseline scores. Underlying the pattern of the ATE on outcome test scores is schools' redistribution of their inputs away from low-achieving students toward high-achieving students, as shown in the second and third panels of Figure 5(b). Schools decrease (increase) their effort inputs for students with lower (higher) baseline scores. In addition, schools also change their

Figure 5: The effect of increasing in proficiency standards to maximize average achievement



(a) (b)

tracking regimes such that peer quality decreases (increases) for low-achieving (high-achieving) students. Finally, the bottom panel shows that the ATE on parental effort moves in the opposite direction to school inputs, which mitigates but does not reverse the effects of school input adjustment on student outcomes.

8 Conclusion

We have developed and estimated a model of ability tracking, in which a school's tracking regime, track-specific inputs, parental effort and student achievement are joint equilibrium outcomes. The estimated model fits the data well.

Using the estimated model, we have shown that the effects of tracking are heterogeneous across students with different prior test scores. In response to the exogenous changes in peer composition under a ban on tracking, schools on average increase their effort inputs for all students. Parents of low-ability students, who are more likely to have low prior scores, decrease their effort while those of high-ability students increase theirs. As a result, banning tracking would increase the performance

for students with below-median prior scores by 2.2% sd and decrease performance for those with above-median prior scores by 4.2% sd.

These seemingly small effects confound the change in peer quality and that in effort inputs, both of which have *significant impacts* on achievement. In fact, ignoring endogenous effort changes would cause one to overstate the gains from banning tracking by over 100% for students with below-median prior scores, and overstate the loss by over 100% for students with above-median prior scores. Our work, therefore, re-emphasizes the importance of taking into account behavioral responses when evaluating policy changes, as argued in Becker and Tomes (1976). It is worth noting that this general principle applies regardless of whether the estimation is done on observational or experimental data.[48]

Because policies that affect schools' tracking decisions will lead to increases of peer quality for some students and decreases for some other students, our study highlights the trade-offs involved in achieving certain educational policy goals. In particular, we have shown that a change in proficiency standards that maximizes average achievement would lead schools to redistribute inputs away from low-ability students toward high-ability ones, in terms of both peer quality and track-specific effort. As a result, the performance of high-ability students increases at the cost of low-ability students.

Our work is promising for future research. Researchers have been expanding the depth and width of datasets, so it is not overly optimistic to think that a dataset containing similar information as the ECLS-K, but at a larger scale per school, will be available in the near future. Our methodology can be easily applied to such a dataset, and, due to its tractability, would remain computationally feasible even when the dataset is large. Moreover, our work also admits potentially interesting extensions for future research. One extension of particular interest would combine studies on the matching between schools and households and this paper into a single framework. This extension would form a more comprehensive view of how peer composition is determined both between and within schools. An even more ambitious project may take into account residential sorting (Epple and Romano (1998), Ferreyra (2007)) or even the responses in housing prices as discussed in Avery and Pathak (2015). Another important extension would allow teachers to

[48] Carrell et al. (2013) provides a good example of this principle in the context of experimental data.

vary in quality and examine how schools match students with teachers, within a tracking framework.

References

Argys, L. M., D. I. Rees and D. J. Brewer, "Detracking America's Schools: Equity at Zero Cost?" *Journal of Policy analysis and Management*, 15(4):623–645, 1996.

Avery, C. and P. A. Pathak, "The Distributional Consequences of Public School Choice," *Working Paper*, 2015.

Becker, G. S. and N. Tomes, "Child Endowments and the Quantity and Quality of Children," *Journal of Political Economy*, 84(4):S143–S162, 1976.

Betts, J., "The Economics of Tracking in Education," *Handbook of the Economics of Education*, 3:341–381, 2011.

Betts, J. and J. Shkolnik, "Key Difficulties in Identifying the Effects of Ability Grouping on Student Achievement," *Economics of Education Review*, 19(1):21–26, 2000a.

Betts, J. R. and J. L. Shkolnik, "The Effects of Ability Grouping on Student Achievement and Resource Allocation in Secondary Schools," *Economics of Education Review*, 19(1):1–15, 2000b.

Blume, L. E., W. A. Brock, S. N. Durlauf and Y. M. Ioannides, "Identification of Social Interactions," in A. B. Jess Benhabib and M. O. Jackson, eds., "Handbook of Social Economics," vol. 1 of *Handbook of Social Economics*, pp. 853 – 964, North-Holland, 2011.

Blume, L. E., S. N. Durlauf et al., "Identifying social interactions: A review," *Methods in Social Epidemiology*, 287, 2006.

Bond, T. N. and K. Lang, "The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results," *Review of Economics and Statistics*, 95(5):1468–1479, 2013.

Bramoullé, Y., H. Djebbari and B. Fortin, "Identification of Peer Effects Through Social Networks," *Journal of Econometrics*, 150(1):41–55, 2009.

Brock, W. A. and S. N. Durlauf, "Discrete Choice with Social Interactions," *The Review of Economic Studies*, 68(2):235–260, 2001.

Burke, M. A. and T. R. Sass, "Classroom Peer Effects and Student Achievement," *Journal of Labor Economics*, 31(1):51–82, 2013.

Carnoy, M. and S. Loeb, "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis*, 24(4):305–331, 2002.

Carrell, S. E., B. I. Sacerdote and J. E. West, "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," *Econometrica*, 81(3):855–882, 2013.

Caucutt, E. M., "Educational Vouchers When There Are Peer Group Effects – Size Matters," *International Economic Review*, 43(1):195–222, 2002.

Cunha, F., J. Heckman and S. Schennach, "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78(3):883–931, 2010.

Das, J., S. Dercon, J. Habyarimana, P. Krishnan, K. Muralidharan and V. Sundararaman, "School Inputs, Household Substitution, and Test Scores," *American Economic Journal: Applied Economics*, 5(2):29–57, 2013.

De Fraja, G., T. Oliveira and L. Zanchi, "Must Try Harder: Evaluating the Role of Effort in Educational Attainment," *The Review of Economics and Statistics*, 92(3):577–597, 2010.

Dee, T. S. and B. Jacob, "The Impact of No Child Left Behind on Student Achievement," *Journal of Policy Analysis and Management*, 30(3):418–446, 2011.

Duflo, E., P. Dupas and M. Kremer, "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101:1739–1774, 2011.

Epple, D., E. Newlon and R. Romano, "Ability Tracking, School Competition, and the Distribution of Educational Benefits," *Journal of Public Economics*, 83(1):1–48, 2002.

Epple, D. and R. Romano, "Competition Between Private and Public Schools, Vouchers, and Peer-Group Effects," *American Economic Review*, 88(1):33–62, 1998.

—, "Peer Effects in Education: A Survey of the Theory and Evidence," in A. B. Jess Benhabib and M. O. Jackson, eds., "Handbook of Social Economics," vol. 1, pp. 1053–1163, North-Holland, 2011.

Ferreyra, M., "Estimating the Effects of Private School Vouchers in Multidistrict Economies," *American Economic Review*, pp. 789–817, 2007.

Ferreyra, M. M. and P. J. Liang, "Information Asymmetry and Equilibrium Monitoring in Education," *Journal of Public Economics*, 96(1):237–254, 2012.

Figlio, D. N. and M. E. Page, "School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?" *Journal of Urban Economics*, 51(3):497–514, 2002.

Fruehwirth, J. C., "Identifying Peer Achievement Spillovers: Implications for Desegregation and the Achievement Gap," *Quantitative Economics*, 4(1):85–124, 2013.

—, "Does the Education of Peers' Mothers and Fathers Matter? Mechanisms of Parental Spillovers in the Classroom," *Working paper*, 2014.

Gamoran, A., "The Variable Effects of High School Tracking," *American Sociological Review*, pp. 812–828, 1992.

Gamoran, A. and M. Berends, "The Effects of Stratification in Secondary Schools: Synthesis of Survey and Ethnographic Research," *Review of Educational Research*, 57(4):415–435, 1987.

Gamoran, A. and M. T. Hallinan, "Tracking Students for Instruction," in "Restructuring Schools," pp. 113–131, Springer, 1995.

Hallinan, M. T., "The Effects of Ability Grouping in Secondary Schools: A Response to Slavin's Best-Evidence Synthesis," *Review of Educational Research*, pp. 501–504, 1990.

—, "The Organization of Students for Instruction in the Middle School," *Sociology of Education*, pp. 114–127, 1992.

—, "Tracking: From Theory to Practice," *Sociology of Education*, pp. 79–84, 1994.

Hanushek, E. and L. Woessmann, "Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries," *The Economic Journal*, 116(510):C63–C76, 2006.

Hanushek, E. A., J. F. Kain, J. M. Markman and S. G. Rivkin, "Does peer ability affect student achievement?" *Journal of Applied Econometrics*, 18(5):527–544, 2003.

Hanushek, E. A. and M. E. Raymond, "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management*, 24(2):297–327, 2005.

Hoffer, T. B., "Middle School Ability Grouping and Student Achievement in Science and Mathematics," *Educational Evaluation and Policy Analysis*, 14(3):205–227, 1992.

Houtenville, A. J. and K. S. Conway, "Parental Effort, School Resources, and Student Achievement," *Journal of Human Resources*, 43(2):437–453, 2008.

Hoxby, C. M. and G. Weingarth, "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects," *NBER Working Paper*, 2005.

Imberman, S. A., A. D. Kugler and B. I. Sacerdote, "Katrina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees," *The American Economic Review*, pp. 2048–2082, 2012.

Jacob, B. A., "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics*, 89(5):761–796, 2005.

Kiss, D., "The Impact of Peer Achievement and Peer Heterogeneity on Own Achievement Growth: Evidence from School Transitions," *Economics of Education Review*, 37:58–65, 2013.

Klein, S. P., L. S. Hamilton, D. F. McCaffrey, B. M. Stecher et al., *What do test scores in Texas tell us?*, Rand Santa Monica, CA, 2000.

Koretz, D. M. and S. I. Barron, *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS).*, ERIC, 1998.

Lavy, V. and A. Schlosser, "Mechanisms and Impacts of Gender Peer Effects at School," *American Economic Journal: Applied Economics*, pp. 1–33, 2011.

Lefgren, L., "Educational Peer Effects and the Chicago Public Schools," *Journal of Urban Economics*, 56(2):169–191, 2004.

Linn, R. L., "Assessments and Accountability," *Educational Researcher*, 29(2):4–16, 2000.

Liu, H., T. A. Mroz and W. Van der Klaauw, "Maternal Employment, Migration, and Child Development," *Journal of Econometrics*, 156(1):212–228, 2010.

Manski, C., "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 60(3):531–542, 1993.

Mehta, N., "Competition in Public School Districts: Charter School Entry, Student Sorting, and School Input Determination," *International Economic Review*, 58(4), 2017.

Mishel, L. R., R. Rothstein, A. B. Krueger, E. A. Hanushek and J. K. Rice, *The Class Size Debate*, Economic Policy Institute, 2002.

Moffitt, R., "Policy Interventions, Low-Level Equilibria, and Social Interactions," *Social Dynamics*, 4:45–82, 2001.

Nechyba, T., "Mobility, Targeting, and Private-School Vouchers," *American Economic Review*, 90(1):130–146, 2000.

Pop-Eleches, C. and M. Urquiola, "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, 103(4):1289–1324, 2013.

Reback, R., J. Rockoff and H. L. Schwartz, "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under No Child Left Behind," *American Economic Journal: Economic Policy*, 6(3):207–241, 2014.

Sacerdote, B., "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" in E. Hanushek, S. Machin and L. Woessmann, eds., "Handbook of the Economics of Education," vol. 3, chap. 4, pp. 249–277, Elsevier, 1st edn., 2011.

Slavin, R., "Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis," *Review of Educational Research*, 60(3):471–499, 1990.

Slavin, R. E., "Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis," *Review of Educational Research*, 57(3):293–336, 1987.

Stinebrickner, R. and T. Stinebrickner, "Time-use and College Outcomes," *Journal of Econometrics*, 2004.

Todd, P. and K. Wolpin, "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The Economic Journal*, 113(485):F3–F33, 2003.

Vigdor, J. and T. Nechyba, *Peer Effects in North Carolina Public Schools*, Schools and the Equal Opportunity Problem?, MIT Press, 2007.

Wong, M., T. D. Cook and P. M. Steiner, "No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series," *Institute for Policy Research (Working paper 09-11)*, 2009.

Appendix

A Functional Forms

A.1 Type Distribution

Denote observable characteristics $x = (x^a, x^p)$, where x^a is the prior test score and x^p includes parent education level and whether or not it is a single-parent household.

Each school has three ability levels $(a_l^s, l = 1, 2, 3)$. Let T_l^s be the l^{th} tercile of $F_{s,a}(\cdot)$, which is the normal distribution approximation of prior test scores of all students in school s, $(\{x_{si}^a\}_i)$. A level a_l^s is defined as the expectation of prior score within the l^{th} tercile in school s computed using $F_{s,a}(\cdot)$, i.e.,

$$
\begin{aligned}
a_1^s &= \int_{-\infty}^{T_1^s} a dF_{s,a}(a|a \leq T_1^s), \\
a_2^s &= \int_{T_1^s}^{T_2^s} a dF_{s,a}(a|T_1^s < a \leq T_2^s), \\
a_3^s &= \int_{T_2^s}^{\infty} a dF_{s,a}(a|T_2^s < a).
\end{aligned}
$$

The distribution of type conditional on x is assumed to take the form

$$
\Pr\left((a_l^s, z)\,|x, s\right) = \Pr\left(a = a_l^s|x^a, s\right) \Pr\left(z|x^p, a_l^s\right),
$$

$$
\begin{aligned}
\Pr\left(a = a_1^s|x^a, s\right) &= 1 - \Phi\left(\frac{x^a - T_1^s}{\sigma_a}\right) \\
\Pr\left(a = a_3^s|x^a, s\right) &= \Phi\left(\frac{x^a - T_2^s}{\sigma_a}\right) \\
\Pr\left(a = a_2^s|x^a, s\right) &= 1 - \Pr\left(a = a_1^s|x^a\right) - \Pr\left(a = a_3^s|x^a\right),
\end{aligned}
$$

where σ_a is a parameter to be estimated. The probability that a parent is of a

high-cost type is given by

$$\Pr\left(z = z_2 | x^p, a_l^s\right) = \Phi\left(\theta_0^c + \theta_1^c a_l^s + \theta_2^c \mathbf{1}\{x_2^p = \text{single parent}\} + \theta_3^c \mathbf{1}\{x_1^p \geq \text{college}\}\right). \tag{6}$$

A.2 Effort measurement system

A.2.1 Parental effort

Observed parental effort $\widetilde{e}^p$ is discrete and ordered in 5 categories, as in Question HEQ.095 in the Spring 6 Parent Questionnaire.

"During this school year, how often did someone help CHILD with his/her reading, language arts or spelling homework? Would you say...

A. Never; B. Less than once a week; C. 1 to 2 times a week; D. 3 to 4 times a week; E. 5 or more times a week; F. REFUSED; G. DON'T KNOW."

Model parental effort is unobserved, but the measurement system below maps e^{p*} and an i.i.d. error term $\zeta^p \sim N(0,1)$ into discrete ordered levels:

$$\widetilde{e}^p = \begin{cases} \text{never} & \Leftrightarrow -\infty < e^{p*} + \zeta^p \leq \kappa_0 \\ < 1 & \Leftrightarrow \kappa_0 < e^{p*} + \zeta^p \leq \kappa_1 \\ \in [1,2] & \Leftrightarrow \kappa_1 < e^{p*} + \zeta^p \leq \kappa_2 \\ \in [3,4] & \Leftrightarrow \kappa_2 < e^{p*} + \zeta^p \leq \kappa_3 \\ \geq 5 & \Leftrightarrow \kappa_3 < e^{p*} + \zeta^p \leq \infty, \end{cases} \tag{7}$$

resulting in $\Pr(\widetilde{e}^p | e^{p*})$.

A.2.2 School effort

School effort is measured in hours per week, according to Question 2 in the Spring 6 Teacher Questionnaire.

"For subjects you teach, about how much time do you expect children to spend on homework in each of the following area (Reading and Language Arts) on a typical evening?

A. I don't teach this subject; B. None; C. 10 min; D. 20 min; E. 30 min; F. More

than 30 min."[49]

The observed effort is given by $\widetilde{e}_j^s = e_j^{s*} + \zeta_j^s$,where $\zeta_j^s \sim N\left(0, \sigma_{\zeta^s}^2\right)$.

B Additional Tables

Table 9: Data: Proficiency cutoffs by Census region

Region name	Proficiency cutoff	Corresponding sample percentile
Northeast	49.69	42.34
Midwest	48.85	34.21
South	43.61	21.41
West	45.69	26.25

Table 10: Data: Prior Test Scores By Track

Track	Mean	Standard deviation	N
1	48.28	11.02	886
2	52.28	8.44	1,227
3	53.99	8.27	563
4	54.99	6.54	113

Table 11: Data: Standardized[1] Prior Test Scores By Track

Track	Mean	Standard deviation	N
1	0.93	0.18	886
2	1.02*	0.15	1,227
3	1.05*	0.14	563
4	1.09*	0.12	113

Note 1: Prior scores are standardized through division by school average prior score.
Note *: Significantly greater than previous standardized track mean (p-value < 0.002).

[49]We treat both E and F as 30 minutes per day.

Table 12: Other Parameter Estimates

Parameter	Estimate	Standard error	
σ_a	8.94	0.55	sd of shock in ability distribution
σ_ϵ	6.72	0.16	sd test score measurement error
σ_{ζ^s}	0.56	0.05	sd school effort measurement error
κ_0	-0.17	0.05	cut-point 1, parent effort measurement
κ_1	0.78	0.03	cut-point 2, parent effort measurement
κ_2	1.98	0.03	cut-point 3, parent effort measurement
κ_3	2.78	0.04	cut-point 4, parent effort measurement

Table 13: Regression of School Intercepts on Household Characteristics and Prior Score

Variable	Coefficient	(Std. Err.)
College	1.491	(0.163)
Single-parent household	-0.633	(0.196)
Prior score	0.097	(0.008)
Intercept	-69.413	(0.430)

Note: "College" means highest education level of parents is college or higher.

Table 14: Summary Statistics: Full Sample vs. Selected Sample (5th Grade)

Variable	Whole Sample			Selected Sample		
	N	Mean	Standard Deviation	N	Mean	Standard Deviation
Sample Size	8,853			2,789		
Parent College	8,047	0.55	0.50	2,789	0.59	0.49
Single Parent	7,953	0.23	0.42	2,789	0.19	0.39
Prior Test Score	8,562	50.17	9.72	2,789	51.47	9.52
Outcome Test Score	8,751	50.15	9.65	2,777	51.68	9.40
Parental Effort	7,788	2.29	1.55	2,703	2.22	1.53
Teacher Effort	2,784	2.03	0.58	533	1.90	0.59
Student Obs. in the School	1,772	5.00	5.54	205	13.61	3.38

Note: "Whole Sample": all fifth graders observed in the ECLS-K.
"Selected Sample": the sample used for our empirical analysis.

Table 15: Summary Statistics: Whole Sample vs. Stayers (Kindergarten)

Variable	All Kindergartners[1a]			Stayers[1b]		
	N	Mean	Standard Deviation	N	Mean	Standard Deviation
Sample Size	16,665			8,494		
Parent College	15,705	0.49	0.50	8,185	0.51	0.50
Single Parent	14,329	0.25	0.43	7,772	0.21	0.41
Prior Test Score	13,734	49.19	9.83	7,039	50.00	9.69
Outcome Test Score	14,579	49.56	9.87	7,831	50.24	9.60

Note 1a: All kindergartners observed in the ECLS-K.
Note 1b: Kindergartners who are observed in the fifth grade whole sample.
Note 2: The statistics are measured in the first survey round, where parental and teacher effort are both unreported.

C Nonlinear Peer Effects

There is recent interest in nonlinear peer effects, where the marginal effect of a change in peer quality is not uniform across students (Sacerdote (2011)). It is important to note that, even in our preferred specification which excludes an interaction term between own ability and peer quality in the technology, the reduced form

mapping between peer quality and achievement is inherently nonlinear (i.e. depends on own ability), because we endogenize school and parental effort decisions.

To see this, consider a student with ability a_i in a track with peer quality q_j. To derive the reduced-form achievement function, substitute her parents' optimal effort e_i^{p*} and optimal track effort e_j^{s*} into (4), and differentiate with respect to peer quality:

$$\frac{\partial Y}{\partial q} = \alpha_3 + \alpha_1 \frac{\partial e_j^{s*}}{\partial q} + \alpha_2 \frac{\partial e_i^{p*}}{\partial q} + \alpha_5 \left(e_i^{p*} \frac{\partial e_j^{s*}}{\partial q} + e_j^{s*} \frac{\partial e_i^{p*}}{\partial q} \right).$$

To ease exposition, suppose $\frac{\partial e_j^{s*}}{\partial q} = 0$, which results in the expression

$$\frac{\partial Y}{\partial q} = \alpha_3 + \frac{\partial e_i^{p*}}{\partial q} \underbrace{\left(\alpha_2 + \alpha_5 e_j^{s*} \right)}_{>0}.$$

Due to curvature of the parents' objective, parents will have different responses to changes in peer quality depending on their children's ability levels, i.e., $\frac{\partial^2 e_i^{p*}}{\partial q \partial a} \neq 0$. This implies that $\frac{\partial^2 Y}{\partial q \partial a} \neq 0$, which is what we set out to show.

D Alternative Production Function Specifications

In each specification, student achievement is governed by a different production function $Y(a, q, q^{cv}, e^s, e^p, \alpha_{0s})$. In particular,

$$\begin{aligned}
&\text{Specification 1:} \quad Y(\cdot) = \alpha_{0s} + a + \alpha_1 e^s + \alpha_2 e^p + \alpha_3 q + \alpha_4 e^s a + \alpha_5 e^s e^p \\
&\text{Specification 2:} \quad Y(\cdot) = \alpha_{0s} + a + \alpha_1 e^s + \alpha_2 e^p + \alpha_3 q + \alpha_4 e^s a + \alpha_5 e^s e^p + \alpha_6 \mathbf{1}\{a > q\} q \\
&\text{Specification 3:} \quad Y(\cdot) = \alpha_{0s} + a + \alpha_1 e^s + \alpha_2 e^p + \alpha_3 q + \alpha_4 e^s a + \alpha_5 e^s e^p + \alpha_7 e^s q^{cv} \\
&\text{Specification 4:} \quad Y(\cdot) = \alpha_{0s} + a + \alpha_1 e^s + \alpha_2 e^p + \alpha_3 q + \alpha_4 e^s a + \alpha_5 e^s e^p + \alpha_6 \mathbf{1}\{a > q\} q + \alpha_7 e^s q^{cv}
\end{aligned}$$

The first specification includes school-specific intercepts. The second specification adds a nonlinear effect of peer quality, through $\alpha_6 \mathbf{1}\{a > q\} q$, allowing the effect of peer quality to differ based on one's own ability. In the terminology of Hoxby and Weingarth (2005), this allows for a "single-crossing" model of peer effects. The third specification adds to the first one, adding an interaction between teacher effort and the track-level coefficient of variation, $\alpha_7 e^s q^{cv}$. In the terminology of Hoxby and Weingarth (2005), this allows for a "boutique" model of peer effects, wherein

Table 16: Summary of Main Results, Across Production Function Specifications

	Specification			
	1	2	3	4
ATE				
All students	-0.005	-0.009	0.001	-0.008
By prior score				
Below-median	0.039	0.024	0.032	0.022
Above-median	-0.055	-0.049	-0.034	-0.042
Marginal effect of 1 sd increase in peer quality				
Ceteris paribus	0.24	0.20	0.19	0.20
Composite	0.13	0.10	0.08	0.09
Bias from ignoring effort adjustment (%)				
Below-median y0	112	137	102	147
Above-median y0	70	85	114	121
log likelihood	-16459.5	-16440.1	-16432.7	-16426.7

Note: Values are expressed in sd of outcome score, unless otherwise noted.

student homogeneity may facilitate designing relevant lesson plans and classroom activities. The fourth specification is the focal specification presented in this paper, and nests specifications 1-3.

Appendix: Data Issues

E Tracking Intensity

This section characterizes the extent to which tracks within a school differ from one another, i.e., tracking intensity, and how tracking intensity differs between schools. To do this, for each school that tracks students we computed the mean prior test score for each track (track mean), and then computed the range in track means within each school. We find that the difference between the mean prior score in the highest versus lowest ability tracks varies considerably between schools: the 25th percentile difference is 0.29 (sample) sd of prior test scores, the median and mean differences are 0.98 and 1.02 sd, respectively, and the 75th percentile difference is 1.59 sd. That is, conditional on it being implemented, tracking still means different things across schools. Our model is designed to be flexible enough to accommodate this type of heterogeneity.

F Further Data Analyses and Robustness Checks

In this section, we examine potential misspecification of tracking regimes, by testing the assumption of track monotonicity and examining between- versus within-track variation. Before we start, we would like to highlight that differences between tracks can be small or large; these differences depend on the tracking decisions made by schools, which are fully incorporated in the model.

F.1 Track Monotonicity

F.1.1 Prior Scores by Track

To start, we examine mean prior scores, which are measures of peer quality, by track. To confirm that prior scores are not lower in "higher" tracks, we regress students' prior scores on their track identity separately for each school, and test whether each track had a mean prior score at least as high as the track just below. At the 5% significance level we reject this hypothesis 15 times, using a one-sided test. Out

of the 547 tracks in the sample, 538 came from schools with more than one track, meaning monotonicity is violated in less than 3% of tracks.

For a more detailed analysis, consider, for example, the following regression of prior score on track identity for a school with three tracks:

$$\text{Prior Score}_{si} = \beta_{s1} + \beta_{s2}\mathbf{1}\{i \text{ is in track 2}\} + \beta_{s3}\mathbf{1}\{i \text{ is in track 3}\} + \nu_{si}. \quad (8)$$

We test whether track monotonicity was violated at this school by testing whether $\beta_{s2} = 0$ versus an alternative of $\beta_{s2} \neq 0$ and $\beta_{s3} = 0$ versus an alternative of $\beta_{s3} \neq 0$; a statistically significant violation of monotonicity would be a case where the data rejected the null hypothesis *and* the coefficient was negative, e.g., $\hat{\beta}_{s3}$ was significantly less than zero. We conduct this test by running the regression (8) separately for each school (suitably amended to take into account the number of tracks at the school), which affords 342 pairwise comparisons. There are four cases, corresponding to coefficient sign and significance level (5%). Case 1 consists of 114 comparisons, in which we reject that the higher track had the same mean prior score, in favor of the alternative that it had a higher mean prior score. Case 2 consists of 139 comparisons, where the estimated coefficients are positive, but insignificant. Case 3 consists of 74 comparisons, where the coefficients are insignificant and negative. Case 4 consists of 15 comparisons with statistically significant negative coefficients.

F.1.2 Parental Inputs

We use data on parental effort to further examine the issue of potential misspecification of tracking regimes. First, we show that parents make decisions based on the contents of their children's track (e.g., peer quality), not the ID of the track. This can be seen in Table 17. Column (1) presents results from a regression of parental effort on track ID, where the reference category is the lowest track (track 1). We can see that, as in the descriptive results in Table 3, parents provide less effort in higher tracks; this difference is statistically significant at the 10% level for parents of students in track 3. However, this relationship disappears in Column (2), where we also condition on a peer quality measure (mean prior score), and remains insignificant in Column (3), which also conditions on household characteristics. That is, the sheer number of schools we find to be tracking is not as important as the fact that other inputs (e.g., parental effort) co-move with variation in peer quality, not

tracking labels.

Second, we examine how differences in parental effort between adjacent tracks relate to differences in peer quality; we should not see large differences in parental effort when considering tracks that have very similar peer quality. Our model allows for parents to react to peer quality in a compensatory way, and the magnitude of such effort adjustment depends on the magnitude of the change in peer quality. This is precisely the pattern in the data and what we find in our structural estimation.

In Figure 6, the x-axis denotes the difference in mean score between a track and the one adjacent, just below. A positive value indicates the higher track had a higher mean prior score than the lower one. The y-axis denotes the difference in mean parental effort, where if the higher track has lower effort (as is broadly the case in the data) the difference would be negative, i.e., compensatory behavior. Finally, the color of each point indicates the case, from the previous section, into which the prior test score difference falls: black dots are Case 1, green dots are Case 2, blue dots are Case 3, and red dots are Case 4. Notably, both within each case and across the four cases, the difference in parental effort is smaller when track quality differences are smaller. Moreover, in terms of the relationship in parental effort inputs and measure of peer quality, there does not seem to be anything qualitatively different across schools in the four different cases. This leads us to believe Cases 3 and 4 are unlikely to affect our results substantially. To confirm this, we conduct additional checks below.

F.1.3 Robustness Check

In this section we show that Case 3 and/or Case 4 are unlikely to affect our results in any significant way.

Descriptive Statistics We first show how the exclusion of schools featuring of either Cases 3 or 4 does not appreciably affect descriptive statistics. Table 18 provides descriptive statistics for the baseline sample (Column (1)), sample excluding schools involved with Case 4 (Column (2)), and the sample excluding schools involved with either Case 3 or Case 4 (Column (3)). We can see that the distributions of prior scores, outcome scores, and peer quality measure are all quite similar across the three samples, as are the correlation between prior score and peer quality measure

Table 17: Regressions of parental effort on track ID, peer quality measure, and household characteristics

	(1)	(2)	(3)
$\mathbf{1}\{\text{track ID} = 2\}$	-0.0234	0.0490	0.0437
	(-0.31)	(0.65)	(0.58)
$\mathbf{1}\{\text{track ID} = 3\}$	-0.174	-0.0527	-0.0544
	(-1.91)	(-0.57)	(-0.59)
$\mathbf{1}\{\text{track ID} = 4\}$	-0.173	-0.0382	-0.0359
	(-1.05)	(-0.23)	(-0.22)
Peer quality		-0.0284***	-0.0250***
		(-6.26)	(-5.26)
Single parent hh			0.149
			(1.77)
Parent college			-0.121
			(-1.78)
Constant	2.216***	3.645***	3.517***
	(37.87)	(15.46)	(14.60)

Note: t statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

and household characteristics.

Re-estimation To ensure that potential mis-specification of tracking regime is not driving our results, we re-estimate the model twice. First, we excluded schools involved in Case 4. Table 19 shows the re-estimated parameters in Column (2), which are very similar to the original ones, presented in Column (1).

Second, we re-estimated the model excluding any school involved with Cases 3 and/or 4; the re-estimated parameters are presented in Column (3). Most of the parameters are similar to the original ones. The most noticeable differences include: the weight on being above the proficiency standard in the school's objective is about half its baseline value, and the cost of having four tracks is higher, likely reflecting the fact that this subsample has fewer four-track schools.

Figure 6: Adjacent-track difference in mean prior score (x-axis) and parental effort (y-axis)

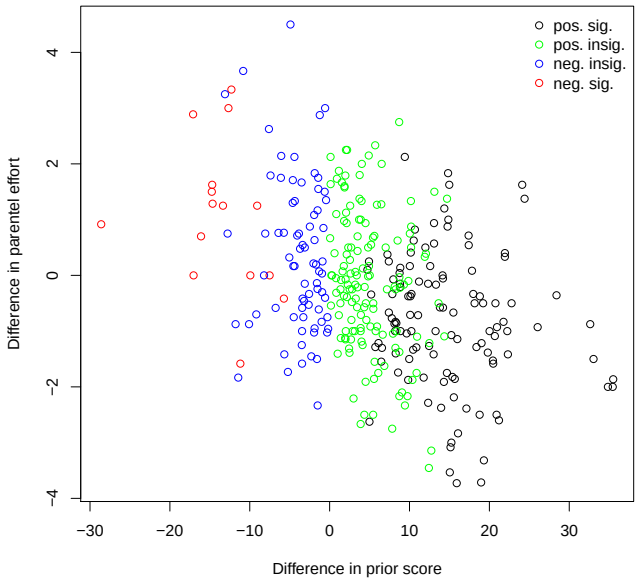


Table 18: Descriptive statistics for baseline sample, sample excluding Case 4, and sample excluding Cases 3 and/or 4

	Baseline (1)	Excl. Case 4 (2)	Excl. Cases 3/4 (3)
Prior score 1st quartile	44.97	45.06	44.82
Prior score mean	51.46	51.63	51.35
Prior score 3rd quartile	57.69	57.81	57.62
Outcome score 1st quartile	45.04	45.25	44.82
Outcome score mean	51.68	51.81	51.62
Outcome score 3rd quartile	57.29	57.41	57.27
Peer quality 1st quartile	47.97	48.23	47.37
Peer quality mean	51.46	51.63	51.35
Peer quality 3rd quartile	55.42	55.52	55.23
Cor(Prior score, peer quality)	0.6936	0.6884	0.6873
Frac. single parent	0.1915	0.1932	0.2118
Frac. college	0.5887	0.5966	0.5612
Number of schools	205	190	121
Number of households	2789	2593	1634

Table 19: Parameter estimates using baseline and restricted samples

	Baseline (1)	Excl. Case 3 (2)	Excl. Cases 3/4 (3)
Production technology			
α_1	9.27	9.27	9.31
α_2	17.68	17.69	16.86
α_3	0.28	0.28	0.20
α_4	-0.05	-0.05	-0.04
α_5	1.71	1.70	1.73
α_6	0.05	0.05	0.09
α_7	-5.35	-4.64	-6.86
Parent objective and types			
c_2^p	0.10	0.10	0.10
z_{c1}	0.08	0.07	0.02
z_{c2}	0.11	0.11	0.05
θ_0^c	-44.02	-44.04	-48.05
θ_1^c	0.85	0.85	0.88
θ_2^c	-10.74	-10.94	-9.58
θ_3^c	3.62	3.57	0.26
School objective			
ω_1	2.45	2.35	0.93
c_2^s	2.06	2.07	2.16
γ_2	-1.03	-1.09	-1.14
γ_3	-0.73	-0.70	-0.59
γ_4	0.41	0.50	1.74
Shocks			
σ_a	8.94	8.39	6.39
σ_ϵ	6.72	6.60	5.98
σ_{ζ^s}	0.56	0.57	0.58
κ_0	-0.17	-0.17	0.01
κ_1	0.78	0.78	0.97
κ_2	1.98	1.98	2.13
κ_3	2.78	2.79	2.92

Counterfactual Experiment We re-run our counterfactual tracking ban, first excluding schools involved in Case 4, and then excluding schools involved in either Case 3 and/or Case 4, both under our original parameter estimates. We also re-run the experiment using the parameter estimates (and observations) from the restricted samples excluding Case 3 and excluding Cases 3 and/or 4. Table 20 presents results from the tracking ban, by prior score. The first three columns show that the effect of banning tracking is virtually identical across the three sample, under the baseline parameter estimates. For example, the effect of banning tracking on students with below-median prior scores would be a 2.23% sd increase in achievement in the baseline (Column (1)), 2.24% sd increase when excluding Case 4 schools (Column (2)), and 2.44% sd when excluding both Case 3 and Case 4 schools (Column (3)). Decreases in achievement for students with above-median prior scores are very similar across the three subsamples.

The results are also very similar under the re-estimated parameters for the subsample excluding any school involved with Case 3 (Column (4)), which corresponds to Column (2) in Table 19: a tracking ban would increase the achievement of below-median prior score students by 2.26% sd and decrease the achievement of students with above-median prior scores by 4.41%. Finally, Column (5) presents the mean change in achievement by prior score, using re-estimated parameters for the subsample excluding any school involved with Cases 3 and/or 4, i.e., Column (3) in Table 19. We can see here that a tracking ban would result in a smaller gain for students with below-median prior scores (of 1.74% sd) and a larger loss for students with above-median prior scores (of 7.08% sd).

Table 20: Effect of tracking ban on achievement (sd) by sample and parameter estimates

	Baseline Estimates			Re-estimated	Re-estimated
	All schools	Excl. Case 4	Excl. Cases 3/4	Excl. Case 4	Excl. Cases 3/4
Prior score:	(1)	(2)	(3)	(4)	(5)
Below-median	0.0223	0.0224	0.0244	0.0226	0.0174
Above-median	-0.0420	-0.0414	-0.0445	-0.0441	-0.0708

F.2 Between- vs. Within-track Variation

The model was designed to allow different schools to have different tracking outcomes. Moreover, though in many schools it may be the case that between-track variation in prior scores is larger than within-track variation, the model can also accommodate the opposite pattern. For example, consider a school with three tracks, where Track 1 has low- and middle-ability students, Track 2 has middle-ability students, and Track 3 has middle- and high-ability students. It would be very likely that between-track variation would be smaller than within-track variation in this case.

It is still useful to examine within- versus between-track variation. To get a sense of how much between-track variation there was within each school, we first computed the mean prior test score for each track in each school and then computed the variance of these between-track scores at the school level. We then computed the average of this school-level variance in track-level prior scores, which is 70.36. To get a sense of the extent of within-track variation within schools, we computed the variance of the prior test score in each track-school combination, and then averaged this over all tracks and schools, returning 52.56. This means that between-track variance is 34% higher than within-track variance.

We can also narrow our focus to the case of mixed versus average tracks: the average within-track variance of prior test scores is 54.45, which is very similar to the number for all track identities. The average between-track variance is now 42.38, which is lower than before. This makes sense, as mixed and average tracks are the most similar to each other in the data. Despite this, there is still a significant difference in mean prior scores between mixed and average tracks in the same school.

G Caveats Arising From Sample Attrition

Although the attrition is largely random, as shown in Table 15, the slight non-randomness leads us to make the following caveats. Case 1: If the non-random attrition reflects the sorting of students across different schools, our sample will represent (hence, our results will apply to) only a subset of schools in the U.S.: those with students from relatively better family backgrounds. In this case, our parameter estimates will not be biased. Case 2: If the attrition reflects higher frequency of

school switches among a subgroup of students, and if the out-flow of such students from one school is associated with an in-flow of similar students into the same school, then students with better family backgrounds will be over-represented in our sample within a school. This may bias our parameter estimates. Given that the sorting of students into different schools has been widely documented in the literature,[50] we feel that concerns arising from Case 1 may be more relevant, i.e., our results may not apply to all schools in the U.S., but a subset of them.

[50]See, for example, Caucutt (2002) and Epple and Romano (1998).