Who Benefits from a Smaller Honors Track?*

Zachary Szlendak
University of Colorado-Boulder

Richard Mansfield
University of Colorado-Boulder

April 30, 2021

Abstract

The vast majority of high school courses in the U.S. separate classrooms into standard and honors tracks. This paper characterizes the efficiency and distributional impact of changing the share of students enrolling in the honors track. We first introduce a model of tracking in which students choose their track for each course, but schools can adjust an array of incentives that implicitly govern the enrollment share of the honors track. We show that determining the administrator's optimal choice of honors track size requires knowledge of a set of treatment effect functions capturing the impact of alternative honors enrollment shares on different parts of the distribution of student predicted performance. We then use rich administrative data from North Carolina public high schools to estimate these treatment effect functions by quintile of predicted performance. Across a wide variety of model specifications and alternative pareto weights over achievement gains for different quintiles, we find that the optimal honors track contains 20% to 30% of a course's students. Furthermore, decreasing the size of honors tracks with more than 35% of students would yield a Pareto improvement across predicted performance quintiles. If all North Carolina high schools adopted the optimal honors program size, we estimate that their students would gain an average of 0.02 test score SDs per course relative to the baseline score distribution, with considerably larger gains for schools making substantial adjustments.

*We thank Terra McKinnish, Brian Cadena, Francisca Antman, Allison Atteberry, Taylor Jaworski, Corey Woodruff, and Brachel Champion as well as seminar participants at the Institute for Defense Analyses and CNA for helpful comments and discussions. This research uses data the North Carolina Education Research Data Center at Duke University. We acknowledge both the North Carolina Department of Public Instruction for collecting and providing this information. Contact: zachary.szlendak@colorado.edu, richard.mansfield@colorado.edu

1 Introduction

Tracking is the process of separating students by ability in order to customize the level of content that students experience. Archbald and Keleher (2008) estimate that over 80% of high schools in the US offer courses that feature multiple tracks representing different paces and rigor. Several papers examine the achievement effect of the track choices of marginal students (e.g. Smith and Todd, 2001; Card and Giuliano, 2016). A number of others consider the impact of introducing tracking or removing it entirely (e.g. Figlio and Page, 2002; Duflo et al., 2011). Yet among schools that offer both honors and regular versions of courses, there is wide variation both across schools and within schools across courses in the share of students that enroll in the honors track. Motivated by the lack of consensus about the optimal honors track size, this paper considers the school's choice of how selective to make its honors track.

The effects of reducing the size of the honors track are ex-ante ambiguous, depend on the initial size of the honors track, and are likely to vary by the type of student. Expanding access to honors versions of courses allows the marginal students to experience the greater rigor and peer quality of the honors track. However, as more students move into honors, the honors track becomes diluted and the regular track experiences a brain drain, decreasing the average student quality in both tracks. Furthermore, after students self-sort, teachers may then alter the level of instruction to align with the new student composition of each track. Other classroom characteristics, such as teacher assignment and class size, may also be affected as decentralized schools reallocate resources between the tracks, further obfuscating the effects of the expansion on different types of students.

We investigate the distributional impact of alternative choices of honors track size by estimating separate flexible functions by category of student preparedness that map a course's fraction in honors into expected standardized test score performance. To justify and motivate our empirical approach, we also introduce a simple theoretical sorting model of a typical high school environment in which students can self-sort into their chosen track for each course, but an administrator can adjust the costs of doing so to implicitly select a preferred honors track size. The model yields conditions under which the functions we estimate are sufficient to determine the administrator's optimal choice of course-wide enrollment shares in each track.

There are three essential challenges to estimating the impact of changing the intensive margin of honors selectivity. First, like other school policy interventions, the expected per-student achievement impact of changing the size of the honors track is likely to be small relative to all of the other student, teacher, and school inputs that affect test score performance.

Thus, the amount of variation necessary to obtain sufficient power to detect treatment effects from alternative honors track shares is daunting, particularly when there are strong theoretical reasons to expect heterogeneous and non-monotonic effects from increased selectivity. In particular, the onerous sample and specification requirements generally preclude the use of small scale experiments and narrowly defined instrumental variables that would otherwise provide credible identification.

Second, because introducing an honors track or changing its size may involve altering not just the depth with which content is covered but also the scope of the curriculum itself, standardized tests may become misaligned with what students are taught, creating measurement error that is correlated with the change in the honors enrollment share. Third, valid identification of the effect of changing the size of honors is empirically difficult because the honors enrollment share is partially endogenous to school, teacher and student characteristics that affect student achievement. For example, an unobservably better-prepared student population might drive both the share of students in honors and average test score performance.

The North Carolina administrative records we use are particularly well-suited to address all three challenges. The data contain histories of elementary and middle school test scores for the near universe of public high school students from 1995 to 2011. In addition, the data feature statewide course-specific tests in eleven high school courses, of which we focus on six that were consistently offered and for which tracks are easily inferred.[1] By facilitating comparisons across schools, across school cohorts, and across courses within a cohort, these two features ensure that an enormous amount of variation in honors track sizes and contemporaneous achievement can be harnessed to identify heterogeneity in impacts at different margins of selectivity for different student subpopulations.

Furthermore, North Carolina's accountability system provides strong incentives to principals and teachers to adhere to the curriculum tested by the statewide exams regardless of track, including test score-based teacher bonuses and public ratings of schools. Such incentives mitigate concerns about misalignment between the content taught versus tested in each track.

Finally, the data provide rich controls at the school, teacher, family, classroom, and student levels, including parental educational attainment, school size, class size, student demographics, and teacher experience, education, and licensing test performance. These controls collectively capture many of the inputs that jointly drive test score performance and the size of the honors track, thus dramatically reducing the scope for simultaneity and

[1]The courses excluded either have multiple advanced tracks such as honors and Advanced Placement, are generally taken in middle school, or are infrequently tested.

omitted variable biases.

In our baseline specification, we pool the cross-sectional, time series, and cross-course variation in the share of a course's students that choose the honors track, since there are plausible sources of potentially exogenous variation at each level. In particular, phone conversations with staff at several North Carolina schools indicated that different principals and department heads exhibit idiosyncratic beliefs on the optimal size of an honors track or preference weights for relative performance of different student subpopulations. Also, relatively modest changes in cohort size may affect the number of classrooms that must be offered in a course to meet class size objectives. This could change the natural set of honors shares depending on the track of the classroom added or removed from offerings.

We then aggregate to the school-course-year-preparedness quintile level, which sidesteps the selection problems associated with individual students' choices of track that have been the focus of much of the tracking literature. We also restrict the sample to schools with typical distributions of student past performance, so that the regular and honors peer environments associated with a given honors fraction are likely to be similar across schools. We then regress test scores on a cubic function of the fraction of students in honors in the associated school-course-year combination, along with our full set of controls. To account for heterogeneity in impact, separate cubic coefficients are estimated for each quintile of a regression index of student preparedness based on past test scores. Validity of our baseline estimates requires that, conditional on our full set of controls, the variation in the share of a course's students that chooses the honors track is unrelated to other unobserved school, teacher, and student inputs that may affect test score performance.

To address remaining endogeneity concerns, we employ several alternative specifications that introduce either fixed effects at various levels or instrumental variables in order to isolate different and in some cases mutually exclusive sources of variation in honors track size. We concede that no single specification represents an airtight identification strategy; instead our confidence in the results stems from their consistency across these specifications. In order for spurious correlations to drive our results, separate sources of endogeneity from different levels of variation would have to generate bias functions with the same pattern and similar magnitudes across the interval of honors enrollment shares for the first quintile of our preparedness index, and would then need to agree again on other bias functions with distinct patterns and magnitude for each of the other four quintiles we consider.

We find that students in the first (highest) predicted achievement quintile benefit the most from honors programs that comprise 20-30% of the student body; they enjoy an expected increase of 0.08 test score standard deviations relative to a version of the course without tracking. The second quintile exhibits similar but smaller effects as the first, with an average

test score gain of about 0.05 standard deviations (SDs) for the 20-30% range, but the test score gains for the second quintile decrease at a slower rate when the share of the student body increases past 30%. The third quintile experiences its largest gains from slightly larger honors programs, gaining an average of 0.04 SD when 25-35% of the student body is enrolled in honors. The fourth quintile is relatively unaffected by variation in the size of the honors program, but does exhibit small gains of about 0.025 SDs relative to the absence of tracking when the share of students in honors is between 20 and 30%. The fifth quintile does not exhibit any statistically or economically significant gains from any exclusiveness relative to trackless courses, but is hurt by the existence of honors tracks with more than 35% of the student body in them.

When administrators value the gains of all quintiles equally, honors tracks with 20-30% of student body enrollment maximize the school's average score, with average gains of 0.04 SDs compared to the absence of an honors track. Furthermore, enough schools and courses feature suboptimal honors enrollment shares so that if all schools switched from their current honors program size to the optimal size, we predict that North Carolina high school students would gain an average of 0.02 test score SDs. The 20-30% range for the share of students in honors still maximizes the weighted average performance and delivers sizable gains relative to no tracking even with a compensatory weighting system that weighs the achievement gains of quintiles 1, 2, 3, and 4 at 20%, 40%, 60%, and 80% of those of quintile 5, respectively. For honors shares greater than 30%, the benefit of having more students placed into the honors program seems to be more than offset by the cost of having both the regular and honors track decrease their average student quality and the level of instruction.

Furthermore, since these relatively small per-student gains would be enjoyed by millions of students and thousands of high schools, changing honors track enrollment shares potentially represents a low cost avenue for generating a substantial aggregate gain in student achievement. Using a back of the envelope calculation that assumes that tracking-induced test score gains generate the same impact on earnings potential as Chetty et al. (2014a,b) found for teacher quality-induced test score gains, we estimate that transitioning all North Carolina high schools' current honors enrollment shares to the optimal 20-30% shares for six core courses would yield an aggregate increase in age 28 earnings of $44 million for each cohort.

Our contribution to the tracking literature is to quantify the impacts of changing the intensive margin of honors track selectivity in a context where students self-select into tracks conditional on capacity constraints implicitly set by school administrators. Other papers have evaluated the extensive margin choice of whether to have any tracking, in several cases by exploiting experimental or quasi-experimental variation. These papers generally do not

analyze the size of the honors track when it exists. Some of these papers have found that tracking helps the top students and hurts the bottom students (Betts and Shkolnik, 2000; Hoffer, 1992; Argys et al., 1996; Epple et al., 2002; Fu and Mehta, 2018). Others have found that tracking does not hurt any students (Zimmer, 2003; Figlio and Page, 2002; Duflo et al., 2011; Card and Giuliano, 2016) or has small or insignificant effects (Pischke and Manning, 2006; Lefgren, 2004). Our results suggest that these seemingly contradictory results might potentially be reconcilable if the different papers feature samples of schools with different mixes of honors enrollment shares.

Fu and Mehta (2018) represents the rare paper in this literature that incorporates an explicit role for honors track selectivity. The authors build a structural model that includes an administrator choosing how to assign elementary school students to different tracks. The model permits heterogeneous effects for tracking schemes that vary with the size of each track. However, while their approach permits a broader welfare analysis, it also requires strong assumptions to simultaneously identify the parameters that govern tracking in combination with other preference and technological parameters related to other choices in the model. Furthermore, they focus on elementary schools, and their tracking data are not as rich or reliable as the North Carolina administrative data.[2]

A second strand of the literature considers the effect on an individual student of moving into an honors or gifted track, either using regression discontinuities (Card and Giuliano, 2016) or propensity score matching (Hoffer, 1992; Long et al., 2012; Smith and Todd, 2001). These papers generally find that enrolling in advanced tracks improves test scores for the marginal students they consider. Our estimates combine the effects on the marginal students with the accompanying effects of diluting the honors track and reducing the peer quality in the regular track. Our results suggest that the impact of honors is not limited to just the marginal students, since we find that students whose past test scores strongly suggest they will be inframarginal are still affected by changes in honors track size.

Finally, this paper also contributes to the much larger literature considering peer effects on academic achievement. While our approach does not isolate the contribution of peer effects, such effects are likely to be one of the driving forces for our results. Hanushek et al. (2006) and Lefgren (2004) find that having better peers improved outcomes for students across the ability distribution. Mehta et al. (2019) find that improved peer quality increases academic performance through both cognitive and non-cognitive mechanisms, such as study time. Imberman et al. (2012) also find that all students benefit from higher achieving peers, but their estimates suggest in addition that the highest ability students are the most sensitive

[2]The authors are forced to infer the track based on variation in teachers' self reports of the quality of their students, which may in some cases reflect sampling variation rather than tracking per se.

to the quality of their peers. Our results are consistent with theirs, since we find that top students gain most from small honors programs, where the peer quality is presumably high, and bottom students are relatively unaffected by small honors programs, suggesting that they are fairly insensitive to peer effects from top students. By adding additional assumptions about student assignment, Fu and Mehta (2018) are able to separately identify peer effects, and similarly find that changing the fraction of students in honors induces changes in peer effects which differ by the type of student affected.

The remainder of the paper will be structured as follows. Section 2 presents a theoretical model that governs the administrator's implicit choice of the size and/or selectivity of the honors track. Section 3 then describes the data, Section 4 lays out the empirical approach, and Section 5 presents the results. Section 6 provides several robustness checks, and Section 7 interprets the findings and concludes.

2 Model

In this section we first describe the planner's tracking problem that the school administrator must solve, which clarifies the required decision inputs that this paper seeks to provide. We then introduce a simple education production function and classroom sorting equilibrium in order to derive a methodology for estimating these decision inputs and to elucidate the assumptions this approach requires.

2.1 The Administrator's Problem

Most high schools allow students to choose their tracks for each course they take. Nonetheless, school administrators have a variety of levers within their control that can alter student incentives to enroll in honors. For example, administrators can preallocate a particular share of classrooms and associated time slots to honors that can affect the scheduling convenience of choosing the honors track. They can also adjust the homework loads in each track, set automatic GPA boosts from taking the honors version of a course, and require mandatory meetings with counselors who can encourage students to enroll in the honors track or discourage them from doing so. Given this reality, rather than assume that administrators can determine the complete allocation of students to tracks for each course, we instead assume that they select the cost of enrolling in honors as a means of implicitly choosing the fraction of students in each track.[3] Given this cost, students' and parents' choices determine the

[3]While most of these levers are not observable in the North Carolina administrative data, GPA boosts are an exception. A simple bivariate regression with course, year, and school fixed effects provides suggestive evidence for our assumption: a one point GPA boost that makes a "B" grade in an honors class equivalent to an "A" in a regular class is associated with a highly significant 12 percentage point increase in the share

particular composition of each course's tracks.

While the administrator can adjust these incentives separately for each course and cohort, we first consider the administrator's problem for an unspecified course and year and temporarily suppress any dependence of the inputs on course and year. Let f denote the chosen fraction of students in honors. Let θ_q denote the preference weight that the administrator gives to the performance of subgroup q, and let W_q denote the share of students in subgroup q among the chosen course and cohort. While these subgroups could be arbitrary combinations of predetermined observable student characteristics, in our empirical work we will use quintiles of predicted student performance based on their test score histories. The weights may reflect adminstrator preferences for gains by different types of students, the relative amount of pressure they face from different groups of parents, administrators, or the priorities for academic growth for different observed types generated by local, state, and federal educational objectives (such as those incentivized by No Child Left Behind). Finally, let $E[Y_i(f)]$ and $E[\overline{Y}_q(f)]$ capture the expected test score of student i and the expected mean test score of students in subgroup q, respectively, as a function of the chosen honors fraction f. We assume that the bulk of the information used by the administrator to predict test scores is contained in the subgroup assignment, so that $E[Y_i(f)] \approx E[\overline{Y}_q(f)]$. Then, assuming further that administrators seek to maximize some weighted average of student performance, we can write the administrator's problem as:

$$\max_f \frac{1}{N} \sum_{i=1}^{N} \theta_{q(i)} E[Y_i(f)] \approx \max_f \sum_{q=1}^{Q} W_q \theta_q E[\overline{Y}_q(f)] \tag{1}$$

This formulation suggests that the principal does not need to predict exactly which students will switch track when the chosen honors fraction changes nor the impact on any given individual from switching track or experiencing a more selective track. Rather, the principal only needs to understand how shifting f changes the mean performance of each subgroup via the new classroom sorting equilibrium. This insight motivates our approach of aggregating over individual track choice and comparing mean outcomes of different subgroups under different tracking regimes. In the next subsection we provide assumptions on the sorting equilibrium that justify this simplified approach.

Under the linear objective function (1), the optimal honors fraction only depends on the degree to which alternative fractions shift test scores of various subgroups, rather than the components of subgroups' mean test scores that are invariant to the honors fraction. Thus, it suffices to focus on the "treatment effect functions" $E[\Delta\overline{Y}_q(f)]$ associated with alternative

choosing the honors track. Unfortunately, GPA boosts are not reported by a sufficient number of districts to be used to form instruments.

choices of f:

$$\underset{f}{\text{argmax}} \sum_{q=1}^{Q} W_q \theta_{q(i)} E[\overline{Y}_q(f)] = \underset{f}{\text{argmax}} \sum_{q=1}^{Q} W_q \theta_q E[\Delta \overline{Y}_q(f)] \tag{2}$$

2.2 Test Score Production

Let Y_{istj} capture the standardized test score of student i in course j taken at school s during year t. We model the educational production function as follows:

$$Y_{istj} = d^h_{istj} \tilde{h}(q_{istj}, \epsilon_{istj} | \vec{q}_h, \vec{\epsilon}_h) + d^r_{istj} \tilde{r}(q_{istj}, \epsilon_{istj} | \vec{q}_r, \vec{\epsilon}_r) + X^O_{istj} \beta^O + X^U_{istj} \beta^U + \mu_{istj}. \tag{3}$$

The student's choice of track is represented by the indicator variables d^h_{istj} and d^r_{istj}, with values of 1 signifying enrollment in honors and regular tracks, respectively. Schools that do not offer separate tracks in a given course feature both $d^h_{istj} = 0$ and $d^r_{istj} = 0$. The functions $\tilde{h}(q_{istj}, \epsilon_{istj} | \vec{q}_r, \vec{\epsilon}_r)$ and $\tilde{r}(q_{istj}, \epsilon_{istj} | \vec{q}_r, \vec{\epsilon}_r)$ capture shifts in achievement from taking the honors and regular tracks, respectively. These shifts are functions of the student's own inputs, which are partly predictable based on the student's observable subgroup q_{istj} but also depend on an unobservable idiosyncratic component ϵ_{istj}. ϵ_{istj} captures deviations in expected performance due to, for example, accumulated skills or effort unaccounted for by subgroup. Such deviations vary not just across students but within students across school-course-year combinations. Importantly, the impact of the track choice on achievement also depends on the peer environment within the chosen track, which is reflected in the dependence of the functions $\tilde{h}(*)$ and $\tilde{r}(*)$ on the vectors $(\vec{q}_h, \vec{\epsilon}_h)$ and $(\vec{q}_r, \vec{\epsilon}_r)$ capturing the subgroups and idiosyncratic contributions of other members of the honors and regular tracks. This flexible formulation of track effects acknowledges that students' production in the classroom will be affected by how the material matches with their ability and how the peer environment interacts with their own ability and effort. Track-specific teacher inputs and course rigor are assumed to be functions of the kinds of students selecting into the track in a given school-course-year, and thus are implicitly captured by the functions $\tilde{h}(*)$ and $\tilde{r}(*)$.

X^O_{istj} and X^U_{istj} capture other observed and unobserved student, school, or course inputs, respectively, that affect i's learning, while μ_{istj} captures measurement error that causes the test score to fail to perfectly reflect the student's learning in the chosen course. Importantly, by imposing that these inputs are additively separable from the inputs that enter the track-specific functions $h(*)$ and $r(*)$, we have assumed they have the same impact on test scores regardless of track. This implicitly requires that the standardized tests used to assess knowledge in each course do not depend on the track chosen, which is true in the

North Carolina context we consider.[4] While somewhat restrictive, the additive separability assumption implies that these inputs are irrelevant to the administrator's tracking problem. Thus, we can rewrite achievement in terms of the difference between performance in the chosen track and performance in a pooled version of the course with no tracks:

$$\Delta Y_{istj} = d^h_{istj} h(q_{istj}, \epsilon_{istj} | \vec{q}_h, \vec{\epsilon}_h) + d^r_{istj} r(q_{istj}, \epsilon_{istj} | \vec{q}_r, \vec{\epsilon}_r). \tag{4}$$

where $h(q_{istj}, \epsilon_{istj} | \vec{q}_h, \vec{\epsilon}_h)$ and $r(q_{istj}, \epsilon_{istj} | \vec{q}_r, \vec{\epsilon}_r)$ now capture the contribution of honors and regular tracks, respectively, compared to a trackless environment. Recasting achievement production this way facilitates a focus on the interactions between the student and peer characteristics that are likely to be of primary importance. Note that this formulation is nonetheless less restrictive than many linear specifications in the literature, since it allows the impact of observed and unobserved student ability components q and ϵ to depend on each other and on the choice of track.

2.3 A Simple Model of Student Track Choice

Now consider the student's choice of honors vs. regular track in a course that features only these two tracks. Suppose that each student chooses the track that maximizes his or her test score net of track-specific effort costs, scheduling opportunity costs, and GPA boosts. Let c_{istj} capture the difference in student i's idiosyncratic composite cost (measured in test-score utility equivalents) of joining the honors track h relative to the regular track r at school s at time t in course j. Next, let α_{stj} capture a component of the composite cost difference that is common to all students in (s, t, j). Importantly, assume that the administrator has the ability to shift α_{stj} by any arbitrary amount by adjusting the relative GPA boost or homework load in the honors track.

The student's track choice can thus be written as:

$$d^h_{istj} = \begin{cases} 1, & \text{if } \underbrace{h(q_{istj}, \epsilon_{istj} | \vec{q}_h, \vec{\epsilon}_h) - r(q_{istj}, \epsilon_{istj} | \vec{q}_r, \vec{\epsilon}_r)}_{\text{Difference in academic gains}} \underbrace{-c_{istj} - \alpha_{stj}}_{\substack{\text{Effort, convenience,} \\ \text{and grade cost}}} > 0 \\ 0, & \text{otherwise} \end{cases}$$

Next, let $g_{stj}(\epsilon, c | q)$ denote the cohort's joint conditional distribution of students' unobserved ability components and idiosyncratic effort/scheduling costs for any given ability group q. To simplify notation, we assume that the school cohorts in consideration are large enough and the ability groups are few enough to approximate $g_{stj}(\epsilon, c | q)$ for each q with a

[4] Furthermore, administrator, parent, and student preferences for high scores help ensure that the curricula for the two tracks do not diverge too far from one another.

continuous joint density. Then we can define $\alpha^*_{stj}(f)$ as the threshold common cost component α^*_{stj} that causes a fraction f of students in the chosen school-year-course to choose the honors track. Specifically, $\alpha^*_{stj}(f)$ is implicitly defined as the solution to the following equation:[5]

$$\sum_q W_{stq} \iint d^h_{istj}(\alpha_{stj}, q, \epsilon, c) g_{stj}(\epsilon, c|q) d\epsilon dc = f. \tag{5}$$

Next, we assume that the composition of students across schools, years, and courses is very similar among a large subset $\mathcal{S}$ of school-year-course combinations:

Assumption 1. $g_{stj}(\epsilon, c|q) \approx g(\epsilon, c|q)\ \forall q\ \forall (s,t,j) \in \mathcal{S}$ *and* $W_{stjq} \approx W_q\ \forall\ (s,t,j) \in \mathcal{S}$

Under Assumption 1, as courses become large the threshold cost function $\alpha^*_{stj}(f)$ becomes common among sufficiently similar schools and course-year combinations within schools: $\alpha_{stj}(f) \approx \alpha_f$ for all $(s,t,j) \in \mathcal{S}$. Furthermore, because the conditional distribution $g(\epsilon, c|d^h, q)$ also becomes common, the vectors of track-specific peers $(\vec{q_r}, \vec{\epsilon_r})$ and $(\vec{q_h}, \vec{\epsilon_h})$ also depend only on f (through $\alpha^*(f)$) rather than separately on s, t, or j. This in turn implies that $h(q_{istj}, \epsilon_{istj}|\vec{q_h}, \vec{\epsilon_h}) \approx h(q_{istj}, \epsilon_{istj}|f_{stj})$ and $r(q_{istj}, \epsilon_{istj}|\vec{q_r}, \vec{\epsilon_r}) \approx r(q_{istj}, \epsilon_{istj}|f_{stj})$. It also implies that the subgroup-specific probability of choosing honors depends only on f:

$$P(d^h = 1|q_{istj} = q, f) = \iint d^h(\alpha^*(f), \epsilon, c, q) g(\epsilon, c|q_{istj} = q) d\epsilon dc \tag{6}$$

Thus, the implicit choice of f by the administrator (through $\alpha^*(f)$) can serve as a sufficient statistic for the peer composition of both the honors and regular tracks in all school-year-course combinations where this common joint distribution of ability and costs represents a sufficiently close approximation. Essentially, this assumption rules out heterogeneous treatments across schools or courses for the same honors fraction, so that differences in achievement distributions across schools or courses featuring different honors fractions can be interpreted as (possibly heterogeneous) treatment effects. In our empirical work, we attempt to make this approximation plausible by removing schools from our sample whose students exhibit a distribution of past performance on state exams that is too far from the state norm.

However, even if the distributions $g(\epsilon, c|q)$ are roughly common among schools, they may not be known by any school administrator, since both ϵ and c are unobserved for

[5]Note that since d^h_{istj} depends on α_{stj} both directly and indirectly through the peer vectors $\vec{q}_h(\alpha_{stj}), \vec{\epsilon}_h(\alpha_{stj})$, we must assume that the track-specific achievement functions $h(*)$ and $r(*)$ are sufficiently insensitive to small changes in peer composition that the fraction choosing honors is monotonically and smoothly decreasing in α_{stj} for each q and spans a large range of fractions for feasible administrator choices of α_{stj}. This ensures that there exist unique solutions to equation 5 for a wide range of f values.

each student. Thus, any given principal will have a difficult time inferring both $g(\epsilon, c|q)$ and the track-specific achievement functions $h(q, \epsilon|f)$ and $f(q, \epsilon|f)$ from data on student performance.

Note, though, that the administrator's problem (1) only requires as inputs $E[\Delta\overline{Y}_q(f)]$, the subgroup-specific mean test score performance gains as functions of the honors fraction f. Thus, we can exploit the fact that $E[\Delta\overline{Y}_q(f)]$ can be written as a simple weighted average of the expected track-specific performance of the subsets of group q that sort into the honors and regular tracks, respectively:

$$E[\Delta\overline{Y}_q(f)] = P(d^h = 1|q, f)E[h(q, \epsilon|f)|d^h = 1] + P(d^r = 1|q, f)E[r(q, \epsilon|f)|d^r = 1] \quad (7)$$

Since the conditional expectation functions $E[h(q, \epsilon|f)|d^h = 1]$ and $E[r(q, \epsilon|f)|d^r = 1]$ in (7) depend only on $g(\epsilon, c|q)$ and $d^h(\alpha^*(f), \epsilon, c, q)$, $h(q, \epsilon|f)$, and $r(q, \epsilon|f)$, which are themselves determined by f through $\alpha^*(f)$, $E[\Delta\overline{Y}_q(f)]$ only depends on the school, course, and year through the administrator's choice of f.[6] Since the objects $E[h(q, \epsilon|f)|d^h = 1]$ and $E[r(q, \epsilon|f)|d^r = 1]$ are means of performance among selected samples of students sorting into each track (partly on the basis of unobserved ability ϵ), they are not objects of interest in their own right, and they do not allow the recovery of the full structural functions $h(q, \epsilon|f)$ $r(q, \epsilon|f)$ without much stronger assumptions on either $h(*)$ and $r(*)$ or $g(\epsilon, c|q)$. However, the above progression makes clear that as long as $g(\epsilon, c|q)$ and W_q are roughly stable for each q across courses and time, identification of the structural functions is unnecessary to solve the administrator's problem.

Essentially, one can simply aggregate over the student-level choice of track, utilizing the fact that every student must choose some track, and compare mean outcomes of students in the same subgroup across schools, cohorts, or courses featuring different administrator choices of f to identify the conditional expectation functions $E[\overline{Y}_q(f)]$ for each subgroup q. Importantly, these functions capture not only the achievement gains or losses from students who have their track choice changed through changes to α_f^* but also how changing f alters the peer effects and level of instruction experienced by other members of the subpopulation.

[6] $E[h(q, \epsilon|f)|d^h = 1]$ and $E[r(q, \epsilon|f)|d^r = 1]$ are defined by:

$$E[h(q, \epsilon|f)|d^h = 1] = \frac{\iint d^h(\alpha^*(f), \epsilon, c, q)h(q, \epsilon|f)g(\epsilon, c|q)d\epsilon dc}{\iint d^h(\alpha^*(f), \epsilon, c, q)g(\epsilon, c|q)d\epsilon dc} \text{ and} \quad (8)$$

$$E[r(q, \epsilon|f)|d^r = 1] = \frac{\iint d^r(\alpha^*(f), \epsilon, c, q)r(q, \epsilon|f)g(\epsilon, c|q)d\epsilon dc}{\iint d^r(\alpha^*(f), \epsilon, c, q)g(\epsilon, c|q)d\epsilon dc} \quad (9)$$

3 Data & Background

We use administrative data provided by the North Carolina Department of Public Instruction for all public schools between 1995 and 2011. These data and their surrounding institutional context have several important features that make it suitable for our analysis.

First, the track associated with each high school classroom is reported for each course, both by school administrators at the beginning of the year and directly by students during assessments at the end of the year. Such dual reporting provides confidence that track is being measured correctly.[7]

Second, North Carolina required statewide standardized end-of course exams as part of 11 distinct high school courses during our sample period. Importantly, because the same exams were administered to all schools and all tracks within a school, these test scores represent a common metric by which to compare schools that choose different shares of honors enrollment.

Of course, drawing valid inferences about relative student learning using test scores from different tracking regimes requires that the alignment between the curriculum and the test content does not systematically vary by track. For example, one might be concerned that much of what is taught in the honors version of the course is not tested on the state exam. However, the North Carolina ABC accountability system that was in place throughout the sample period provides strong incentives for teachers teaching tested courses to adhere to the state curriculum. First, schools are rated publicly based on student test score growth on these exams, and underperforming schools are at risk of sanctions and even closure, so that principals risk their reputations and even their jobs when their teachers do not teach what is tested (Ahn and Vigdor, 2014). Second, teachers at underperforming schools are also at risk of losing their jobs, while teachers at high performing schools are eligible for annual salary bonuses of $1,000 to $1,500 (Vigdor et al., 2008). Finally, student performance on these exams contributes a state-mandated minimum of 25% of the student's course grade, so students have an incentive to study the tested material and parents have an incentive to ensure that teachers adhere to the curriculum regardless of track (Zinth, 2012). Hence, in this North Carolina context, the honors track is likely to primarily represent greater depth and difficulty of covered material rather than greater breadth.

We exclude five of the eleven tested courses from our sample due to either a small set of test years (Civics and Economics, Law & Politics), inconsistency in grade level (Algebra 1 is often offered in middle school rather than high school), or the existence of Advancement

[7]Naturally there are occasional discrepancies due to students misreporting the track of their classroom or students changing track during the academic year. In such cases we use the school-reported track of their classroom in our analysis, but our results are robust to dropping observations featuring discrepancies.

Placement classrooms, discussed further below (US History and Physics). Thus, our sample consists of standardized scores from the following six courses: Algebra 2, Biology, Chemistry, English 1, Geometry, and Physical Science. Appendix Figure A.1, which displays the 2006 statewide distributions of student scores for our final sample from each of the six remaining courses, reveals no evidence of any floor or ceiling effects.[8]

Table 2 examines the tracking options available in each course for all school-year-courses with at least 30 student observations. There exists an honors program in most school-year-course combinations, but remedial programs are rare. Furthermore, the remedial track generally accounts for a very small portion of the student body when it exists (see Figure A.2). Given insufficient power to detect the impact of alternative remedial track sizes, we control for the share of students in remedial classrooms (interacted with quintile of predicted performance), but do not estimate a separate treatment effect function for the remedial track. Because most of the courses tested by state standardized exams tend to be offered in 9th and 10th grade, they do not feature an advanced placement (AP) version of the curriculum. The two exceptions are Physics and U.S. History. We drop these courses from our sample, since we fear that teachers in AP classrooms in these courses may adapt their curricula to align more with the AP exam than the North Carolina end-of-course exam, making the latter exam a less accurate measure of learning.[9] Thus, we focus attention on regular and honors tracks, and use the share of students enrolled in the honors track in a given school-course-year as our main independent variable of interest, in alignment with the honors fraction f in Section 2 above.

Third, the large number of schools, cohorts, and students contained in the North Carolina data ensures that sufficient identifying variation exists to provide properly powered tests of the impact of alternative levels of honors enrollment shares on student performance across the ability distribution. While tracking policy is important because it affects the entire student population in every course, its test score impact per student-course is likely to be relatively small, since much of the variation in student performance is driven by student- and parent-specific factors beyond the school's control. A lack of power has heretofore forced researchers to focus on the extensive margin of whether to offer any tracking rather than the intensive margin of honors selectivity.

Finally, the North Carolina administrative data offers a wide array of observed control

[8]More years are available by request from the authors. No course-year in our sample exhibits bunching around the upper or lower limit of the score range.

[9]We also drop school-year-course combinations featuring classrooms adhering to international baccalaureate standards for the same reason. This proves to be inconsequential, since most schools that offer the IB curriculum are too high-achieving to satisfy our other sample restriction described below that their associated courses plausibly satisfy Assumption 1.

variables at the school, teacher, classroom, and student levels. As emphasized in the following section, such rich controls are critical for addressing omitted variable bias stemming from correlations between the honors share and other school, teacher, and student inputs that contribute to test scores. Of particular note are histories of students' standardized test scores during grades 7 and 8 in math, English and (for some cohorts) science. These histories capture differential student preparedness across schools and cohorts that might both influence principal's decisions about honors track size and predict future student performance.

3.1 Assignment to Preparedness Quintiles and Restricting the Sample of Schools

The test score histories also provide a basis for assigning students to the observed preparedness types that are necessary for providing a holistic assessment of the impact of alternative choices of honors track size. Specifically, we assign each student to a predicted quintile in the statewide performance distribution (with quintile 1 denoting the highest predicted performance) based on the distribution of students' regression indices from a regression of test scores in the sampled high school subjects on grade 7 and 8 English and math scores. We allow the coefficients on these past scores to be course-specific, so that the same student may be assigned to different quintiles for different courses if their past performance indicates different relative strengths in the skills required by these courses.[10] For the sake of brevity, henceforth we refer to these statewide predicted quintiles merely as quintiles, and will be explicit on the occasion in which within-school student rankings are instead used as the basis for assignment to a quintile.

Recall that formal justification for using the fraction in honors as a sufficient statistic for peer environment in each track invoked Assumption 1, which required each school-year-course combination to feature the same joint distribution of abilities (observed and unobserved) and effort costs among students. Clearly this condition will not be satisfied exactly; however, our method only requires that these joint distributions are sufficiently similar across schools and particularly across cohorts and courses within schools so that peer environments would be comparable if honors fractions were equalized. More specifically, we require that comparisons between such units featuring exogenously different honors fractions are informative about how each unit's achievement distribution would change if it were to adjust its own honors fraction.

[10] Specifically, we estimate $Y_{istj} = English7_{istj}\beta_1 j + math7_{istj}\beta_2 j + English8_{istj}\beta_3 j + math8_{istj}\beta_4 j + \epsilon_{istj}$. We then assign course-specific quintiles based on the distribution of $PredictedScore_{istj} = English7_{istj}\hat{\beta}_1 j + math7_{istj}\hat{\beta}_2 j + English8_{istj}\hat{\beta}_3 j + math8_{istj}\hat{\beta}_4 j$. Results are robust to the inclusion of science test scores; however, science scores are only available for a small number of years.

However, a less selective honors track may result in a considerably different ability distribution among students choosing the track at an extremely privileged school relative to a school with few resources and struggling families. To gauge the scale of the problem, Appendix Figure A.4 looks at how many quintiles students would need to shift, on average, in order for each schools' distribution of student preparedness quintiles to match the statewide (uniform) distribution of predicted quintiles. While the majority of schools appear to have nearly uniform distributions, there is a substantial right tail of schools with substantially skewed distributions. We restrict the sample to schools which require fewer than 0.5 quintile changes per student to match the uniform distribution, which removes about 30% of the observations from the original sample. Appendix Figure A.5a shows the histograms of the six schools with required per-student quintile changes closest to and less than one half. While this sample restriction ensures plausible comparability among schools, it may limit the external validity of our estimates for schools with very low or very high student past performance. As a robustness check, we also consider a specification where the above metric for the spread of student quality is less than one third. Appendix Figure A.5b displays histograms of the six school-courses where the average number of quintile shifts required for a uniform distribution are closest to a third.[11]

In addition to restrictions placed on the sets of courses and schools, we also drop all high school test scores from students with missing 7th or 8th grade math or English test scores (since they cannot reliably be assigned to a predicted performance quintile), and we drop all courses offered prior to 1999 to ensure that 7th and 8th middle school test score histories exist for nearly all students in the remaining courses in the sample. Finally, we require each school-course-year in the sample to feature at least 30 tested students so that average characteristics and average performance by quintile would be subject to minimal measurement error. After all of these sample restrictions, our baseline sample contains 2,125,883 total test scores from 335 high schools and 12,882 school-year-course combinations.

3.2 Summary Statistics

If Assumption 1 holds, each principal is perfectly informed about $E[\overline{Y}_q(f)]$, and each has the same preference weights θ_q, then each principal's optimal choice of f to solve (2) would be the same, and there would be no identifying variation in f. Appendix Figure A.3 allays this fear by displaying the distribution of honors shares for the six courses in our final sample among school-year-courses with honors tracks. Every subinterval between 0.1 and 0.6 shows frequent use in all six courses, and Chemistry features a nontrivial share of school-year

[11] Note that North Carolina ranks toward the middle of U.S. states for educational performance, suggesting that our results should be externally valid for most schools throughout the U.S. (U.S. News (2019)).

combinations with more than 60% of students in honors. This suggests that administrators either vary in their preference weights θ_q or hold differential beliefs about $E[\overline{Y}_q(f)]$. Given the dearth of convincing evidence from the literature on the particular tradeoffs associated with different honors track sizes, such varied beliefs are not surprising.

Table 1 decomposes the variance in the honors fraction f among school-year-course combinations in our estimation sample. Unconditionally, differences in school means of honors fractions (pooling across courses and years) account for 37.7% of the total variance, while year-specific deviations from the multi-year mean account for another 12.6%, and course-specific deviations from the school-year mean account for the remaining 49.8%. Adding our baseline control variables (described in the next section) removes about 30% of the total variance, but only slightly changes the contributions of the three decomposition components to the residual variance (to 39.1%, 13.4%, and 47.3%, respectively). When evaluating robustness to our baseline specification later in the paper, we consider several specifications that systematically omit subsets of these components.

Table 3 provides means and standard deviations of the controls used in our baseline specification by category of honors enrollment share based on our estimation sample of school-year-course combinations. Relative to school-year-courses without tracking, those with smaller honors tracks ($< 35\%$ of students) have very similar distributions of student demographics, teacher credentials, and parents' education and only slighty superior past achievement. The one major difference is that school-year-course combinations from small schools (low average cohort size) are more likely not to offer an honors track. Relative to both school-course-years without tracking and with smaller honors tracks, those with larger honors tracks ($> 35\%$ of students) tend to have students with somewhat higher past test scores and slightly more educated parents. Overall, though, the distributions of characteristics across these three honors fraction categories overlap considerably, so at first blush it seems that much of the variation in honors fractions is not directly tied to the composition of students or teachers at the school.

Figure 1 plots in blue the average honors enrollment rate for bins of the coursewide honors fraction separately by within-school (rather than statewide) quintile of predicted performance. We also plot in red the enrollment rate one would expect if students were perfectly sorted to tracks based on their relative predicted performance. Perfect sorting on predicted performance would result in a line with a slope of 5 within the interval of honors fraction corresponding to the chosen quintile ([0,.2] for quintile 1, [.2,.4] for quintile 2, and so on) and a flat line with zero slope elsewhere. The final cell in Figure 1 shows the pooled distribution of honors fractions among all school-course-years in the sample.

Students in top quintiles unsurprisingly enroll in honors at much higher rates than stu-

dents in other quintiles. Nonetheless, the plots of observed honors enrollment patterns reveal quite imperfect sorting, suggesting that unobserved ability and heterogeneous effort costs do play an important role in track choice.[12] For example, a course with 20% of students in honors tends to be chosen by only 58% of top quintile students (rather than the predicted 100%) and by 25%, 11%, 5%, and 1% of students in quintiles, 2-5, respectively. Similarly, a course with 60% of students in honors still has 20% of quintile 2 students enrolling in the regular track while 20% of students in quintile 5 enroll in honors. For quintiles 2 and 3 in particular, these unobserved sorting factors play a large role in track choices, as both quintiles have significant honors enrollment rates for all coursewide shares of students in honors.

Figure 2 plots the average contemporaneous test score performance in test score standard deviations from the statewide mean for bins of the share of students in honors, separately by preparedness quintile. Interestingly, if we disregard the very noisy values for shares of honors above 65% that are observed extremely infrequently in the data, we see that regardless of quintile, the average performance is at or near its peak when the share of students in honors is around 40%. However, in order to verify that this finding reflects a true pareto-optimal honors share rather than a spurious correlation between honors fraction and other school and student inputs, we now describe our more rigorous estimation procedure.

4 Empirical Approach

4.1 Baseline Specification

Our primary specification is an aggregated version of the education production function (3) from Section 2.2. Recall from Section 2.3 that the objects of interest, quintile-specific treatment effect functions of the honors fraction ($E[\Delta\overline{Y}_q(f_{stj})]$), are aggregate objects that only vary at the school-year-course-quintile level. Thus, because the control variables X^O_{istj} enter linearly and are assumed to be additively separable from $E[\Delta\overline{Y}_q(f_{stj})]$, we can estimate the parameters of interest in (3) at the school-year-course-quintile level without introducing any bias and with minimal lost efficiency. Furthermore, such aggregation allows us to avoid selection problems from individual track choice. Thus, our primary specifications all take the following form:

$$\overline{Y}_{stjq} = E[\Delta\overline{Y}_q(f_{stj})] + X_{stjq}\beta^X + \Gamma_{stjq}\beta^\Gamma + \omega_{stjq}. \tag{10}$$

[12] Various measures of ranking on observed ability, including shorter or longer performance history on alternative sets of tests, all show high levels of sorting on unobservables.

Each specification implements $E[\Delta\overline{Y}_q(f_{stj})]$ as a set of quintile-specific, flexibly parameterized functions of f_{stj}, the fraction taking the honors version among all students taking course j in school s in year t, with the chosen functional forms varying across specifications. Our baseline specification imposes that $E[\Delta\overline{Y}_q(f_{stj})]$ for each quintile takes the form of a restricted cubic function:

$$E[\Delta\overline{Y}_q(f_{stj})] = \gamma_q^{lin} f_{stj} + \gamma_q^{sq} f_{stj}^2 - (\gamma_q^{lin} + \gamma_q^{sq}) f_{stj}^3 \tag{11}$$

The coefficients in equation (11) restrict the treatment effect to be the same when placing zero students in honors classes and when placing all students in honors classes, since both scenarios arguably represent an absence of tracking.[13] This functional form allows the location and level of the achievement maximum (or minimum) to be determined by the data while still exploiting the efficiency gains from summarizing a function with two parameters. We also present results from other functional forms, including unrestricted cubic and restricted quartic specifications, as robustness checks in Section 6. Importantly, the coefficients $\vec{\gamma}^{lin} = \{\gamma_1^{lin}, ..., \gamma_5^{lin}\}$ and $\vec{\gamma}^{sq} = \{\gamma_1^{sq}, ..., \gamma_5^{sq}\}$ are quintile-specific in order to capture heterogeneous effects among different levels of student preparedness.

X_{stjq} contains a vector of observed school, teacher, and quintile-mean student control variables that in some cases are specific to the course j and/or year t. Γ_{stjq} represents a design matrix or matrices capturing fixed effects for various one- and two-dimensional combinations of (s, t, j, q). Thus, the theoretical object X_{istj}^O from equation (3) is operationalized as $[X_{stjq}, \Gamma_{stjq}] \equiv \overline{X}_{stjq}^O$ in equation (10). $\omega_{stjq} \equiv \overline{X}_{stjq}^U \beta^U + \overline{\mu}_{stjq}$ captures the combined impact of mean unobserved student, teacher, and school inputs and mean test score measurement error at the (s, t, j, q) level.

Our baseline specification pools all the variation in the honors fraction f_{stj} that occurs between schools, between years within schools, and between courses within school-year combinations. We pool partly to generate maximally precise estimates of the parameters $\vec{\gamma}^{lin}$ and $\vec{\gamma}^{sq}$, but also because there are plausible sources of exogenous variation at each level.

For example, smaller schools may not be able to support the multiple number of classrooms per course that tracking requires, and surpassing the cohort size thresholds beyond which additional classrooms can be supported may not otherwise affect student outcomes (beyond simple class size effects for which we include separate controls). Similarly, due to differential parental pressure, personal pedagogical beliefs, or accountability pressure, principals may differentially weigh performance by different quintiles or have incorrect beliefs about the impact of tracking for reasons unrelated to any of the other unobserved inputs

[13] Three of the largest sources of achievement changes from alternative honors enrollment shares are the same when the fraction of students in honors is equal to zero or one: peer effects, allocation of teachers among tracks, and specialized instruction.

affecting their students' performance, leading to exogeneously different chosen honors fractions. Switching costs from new course preparation for certain teachers may cause schools not to track even when other similar schools do so, perhaps because of differences in the past course histories of their teachers.

Exogenous time series variation in the honors fraction occurring within schools include natural idiosyncratic changes in cohort size that require adding or removing classrooms or deterring or encouraging students to take honors to avoid exceeding classroom capacities. It might also include idiosyncratic variation in the past course preps of newly hired teachers. Exogenous between-course variation stems from idiosyncratic pedagogical preferences by department heads or slightly different student demand for different courses due to scheduling conflicts (which can also vary across cohorts).

On the other hand, the variation in honors fractions at each level is likely to contain an endogenous component as well. Schools may be more likely to dedicate a larger share of course capacity to the honors track when they serve well-prepared students, as suggested by the summary statistics above. And student demand for honors in a particular year may exceed administrator expectations when a cohort is particularly able or motivated. Furthermore, in addition to correlations with unobserved components of student composition, unobserved teacher and school inputs can also be correlated with or actively cause changes in the honors share. For example, perhaps the principals most willing to raise standards for students by encouraging the honors track also invest more time and resources in other achievement-raising policies. Or a school that has particularly effective teachers in a given course may wish to reward them by allowing them to teach honors versions more frequently, and thus increases the share of that course's classrooms that offer the honors version.

Unfortunately, observable variables that isolate only the exogenous sources of variation are either not available or yield instruments that are too weak to detect the hypothesized heterogeneity in achievement impacts across the student ability distribution. Since we estimate the model via ordinary least squares, in order for our baseline estimates to be unbiased, unobserved inputs contained in the error term must be uncorrelated with the honors share f_{stj} as well as its square and cube, conditional on the controls X_{stjq} and Γ_{stjq}.[14] Thus, our baseline specification relies heavily on the richness of the North Carolina administrative data to provide a set of powerful controls that absorbs the most plausible sources of endogeneity.

The full list of baseline controls and their sample means and standard deviations by category of honors share are provided in Table 3. To address bias from correlation with student composition, in our baseline specification the vector X_{stjq} contains student ability

[14] Specifically, we assume $E[\omega_{stjq} f_{stj} | X_{stjq}, \Gamma_{stjq}] = 0$, $E[\omega_{stjq} f_{stj}^2 | X_{stjq}, \Gamma_{stjq}] = 0$, and $E[\omega_{stjq} f_{stj}^3 | X_{stjq}, \Gamma_{stjq}] = 0$.

and preparedness measures (mean test scores in grade 7-8 math and English interacted with quintile, share with gifted status), student demographics (race shares), and family socioeconomic indicators (parental education categories). Note that introducing a variety of such measures as aggregate shares at school and school-course-year-quintile levels rather than at the individual-level makes them stronger controls; at such higher levels of aggregation, they implicitly control for unobserved school and cohort characteristics by potentially spanning the common amenity space that lures certain kinds of observably and unobservably superior or inferior students to the school or course within the school (Altonji and Mansfield, 2018).

To address endogeneity from teacher inputs and remaining school inputs beyond those that affect student composition, X_{stjq} also includes proxies for teacher quality (experience, certification scores, degree/license status), and controls for the school's size and Title 1 status. We also control for both the number of classes offered and the mean class size at the (s, t, j) level, important inputs that may sometimes move in tandem with otherwise idiosyncratic changes in honors shares. Thus, we intend for our treatment effect functions to isolate the impact of changing the honors share conditional on class size (i.e. from converting a class from regular to honors track) rather than combining the impact of simultaneous changes in both class size and honors share (i.e. from adding an extra regular track class to the existing roster of classrooms).

All controls are interacted with the full set of course indicator variables to allow differential predictive power in different courses. Finally, in our baseline specification Γ_{stjq} includes a full set of year-course-quintile fixed effects, which removes potential bias from secular changes in statewide course curricula or the relative difficulty of standardized test questions that target different parts of the ability distribution that may be correlated with statewide trends in honors fractions.

While we believe that these controls adequately address a multitude of potential endogeneity problems, we nonetheless consider three alternative specifications to partially address remaining concerns about simultaneity bias or omitted variable bias.

The first alternative specification adds a set of school fixed effects to Γ_{stjq}, so that the parameters of interest are only identified by differential changes in honors fractions and achievement across cohorts and courses within schools. While school fixed effects address concerns stemming from greater honors enrollment shares causing or responding to student sorting among schools, adding these fixed effects also generates noisier estimates, since between school variation accounts for 39.2% of residual identifying variation net of controls in our baseline specification.

The second alternative specification we consider uses the honors share of the previous cohort in the same school-course combination, along with its square, as instruments for the

corresponding contemporaneous share and its square. The exclusion restriction for this IV specification requires that the past share of students in honors affects current test scores only through inertia in the share of honors over time conditional on controls. This IV approach purges estimates of any endogenous honors share response to unobservable changes in cohort quality or teacher staffing within a school. Implicitly, this specification puts greater emphasis on between-school and within-school/across-course variation at the expense of time-series variation.

The third alternative specification uses a similar IV approach, except that the mean honors share among all other contemporaneous courses (and its corresponding square and cube) are used an instruments for the honors share in the chosen course, its square, and its cube. By generating predicted honors fractions that are pooled across other courses, this leave-one-out IV approach removes any endogeneity stemming from higher teacher quality in particular courses driving higher honors fractions. This estimator essentially relies only on between-school and within-school/across cohort variation instead.

While no single one of these alternative specifications is intended to allay all fears about bias in isolation, collectively they can potentially provide considerable reassurance if results are consistent across all of these specifications, since each level of variation is excluded from at least one of these specifications. After all, if substantial endogeneity biases exist, they would need to operate with the same force (relative to the exogenous variation) at each level of variation *and for each quintile of student preparedness* in order to generate such consistency. Put another way, our flexibility in allowing separate cubic functions of the honors fraction for each quintile also provides more opportunities for sizeable endogeneity biases to reveal themselves through distinct results patterns across specifications that magnify or reduce the role these sources of endogeneity play in driving results.

We cluster standard errors at the school level in each specification, both to be conservative and because we expect considerable autocorrelation in errors across course-years from the same school. In addition, each specification weighs observations by the share of the students at the school-year-course that are in each quintile, so that all school-year-courses are weighted equally.[15]

[15] A weighting scheme based on the number of students rather than within-school-year shares would prioritize the efficacy of administrators' actions at large schools over smaller schools. Given that we are interested in providing inputs to principals of all school types, we prefer weighting schools rather than students equally. As per the recommendations of (Solon et al., 2015), specifications are available upon request in which weights proportional to the number of students in the school-year-course quintile combination. Point estimates and standard errors are similar for the different weighting schemes.

5 Results

5.1 Quintile Treatment Effects

The red lines of Figure 3a display predicted values of treatment effects on achievement scaled in standard deviations of standardized test scores for a dense grid of potential honors fractions from our baseline restricted-cubic specification, which pools all sources of residual variation in the honors fraction among school-year-course combinations. Note that all predicted values capture treatment effects for alternative honors fractions relative to an absence of tracking, which has been normalized to zero. Dashed blue lines indicate the upper and lower bounds on 95% pointwise confidence intervals that were created by using the delta method to convert the variance-covariance matrix associated with point estimates for the cubic parameters $\vec{\gamma}$ into confidence intervals for each predicted value along the grid.[16] The bottom right cell in the figure displays the support of the honors share distribution for school-year-courses that feature an honors track. Note that there is limited support among honors programs with shares greater than 65% or between 0 and 15%, so our predicted values in these ranges are primarily driven by our functional form assumptions and should be viewed skeptically. Table 4 provides the predicted values and the associated 95% confidence intervals separately by quintile for several candidate honors fractions that underlie the Figures 3a - 4b for both our baseline and alternative specifications. Appendix Table 1 provides the underlying parameter estimates $\vec{\gamma_q}$ for each quintile q for these specifications.

Starting with quintile 1, we observe that top students benefit significantly from honors programs with fewer than 30% of students in them: when the treatment effect function reaches its estimated peak honors share of .24, they gain an estimated .079 standard deviations in state test score performance relative to the absence of tracking. This is similar to the predicted increase in student achievement associated with switching from a high school teacher of median effectiveness to a 67^{th} percentile teacher (Mansfield, 2015). However, these gains quickly disappear as the honors fraction increases beyond 30%. Since around 80% of quintile 1 students will enroll in honors if it contains at least 30% of their cohort, the sharp decrease in gains as honors becomes more selective is likely due to the dilution in peer quality within the honors track, perhaps combined with smaller and smaller gains from switching track for the remaining marginal students. Imberman et al. (2012) found that high achieving students are especially sensitive to peer effects, potentially justifying why quintile 1 experiences such a pronounced decrease as the share of students in honors is increased.

[16] We chose pointwise confidence intervals rather than confidence bands because we are generally comparing predicted values at particular honors shares against the absence of tracking, rather than evaluating joint hypotheses involving predicted values over a continuous range of honors fractions, such as whether there exists any nonzero honors fraction that makes quintile 2 worse off than the absence of tracking.

Students in quintiles 2 and 3 also particularly benefit from fairly small honors programs. Relative to the absence of tracks, the gains from the existence of an honors track rise until peak gains of about 0.055 SDs and 0.040 SDs, respectively, when 25% and 28% of all students are choosing honors. Interestingly, this peak occurs at a fraction where large shares of students in these quintiles are near the margin of choosing honors: around 50% of quintile 2 students and 25% for quintile 3 students generally enroll in an honors track that serves 30% of the cohort, with these shares continuing to rise significantly as the coursewide share of students in honors increases beyond 30%.

Several competing mechanisms are potentially at play for these quintiles. As the honors track increases from a very small size to a moderate size, students from these quintiles are likely to be the marginal students, and the pedagogy in the honors track is likely becoming better and better aligned with their desired pace. The regular track is beginning to lose high quality peers, but is still likely to be fairly well aligned with the desired pace for quintile 3 students. As honors selectivity continues to fall, however, there are more inframarginal quintile 2 and 3 students already in the honors track who are experiencing dilution, and the median student in the regular track may increasingly require a slower pace than is optimal for the remaining quintile 2 and 3 students.

Decomposing these competing mechanisms to isolate how each incremental expansion of the honors track affects marginal students, inframarginal honors track students, and inframarginal regular track students within each quintile would require strong assumptions on the degree to which unobservable ability vs. scheduling costs is driving students' selection of track. Indeed, the appeal of our approach is that it can provide the policy-relevant inputs for administrators without requiring questionable assumptions about student sorting to tracks. Thus, we do not attempt such a decomposition here.

Quintile 4 students seem to be fairly insensitive to the size of the honors track. Equivalence of a two track menu with a trackless course can only be rejected with 95% confidence for shares of students in honors less than 30%. The point estimate at the peak, which is at an estimated 21.6%, is .029 SDs.

Quintile 5 exhibits only small, statistically insignificant gains from small honors programs, and begins to experience losses relative to a no tracking regime once the honors program grows beyond 30%. The losses achieve statistical significance at the 95% level around an honors share of 50%. These results are consistent with the peer effect literature that has found that lower achieving students are the least sensitive to the positive peer effects from the highest ability students (Imberman et al., 2012; Mehta et al., 2019; Fruehwirth, 2013; Fu and Mehta, 2018). Although having a small honors program decreases the average peer quality for the overwhelming majority of bottom quintile students who do not enroll in honors (over

90% remain in the regular track with a 40% cohort-wide honors share), the compositional changes may be offset by a better paced class. However, perhaps when the honors program grows beyond 40%, the bottom quintile students who do not enroll in honors (still around 80% when the cohort-wide percent in honors is 60%) share the classroom with fewer middle tier students with whom they might otherwise profitably interact.

We next consider the three alternative specifications introduced in Section 4. Estimated treatment effect functions for these specifications are presented in Figures 3b, 4a, and 4b. Recall that the alternative specifications are identified by different subsets of the variation identifying our baseline model. Figure 3b corresponds to the school fixed effects specification that isolates variation in honors policies that either change over time and/or vary by course within a school. The school fixed effect specification yields quite similar shapes and peak locations to the baseline specification across all quintiles, despite removing 39% of the identifying variation. However, the peak gains from optimally sized honors tracks tend to be around .02 SDs smaller across quintiles.

Figure 4a presents results from the first IV specification, which uses lagged course-specific honors shares as instruments for current honors shares. It seeks to remove cohort-specific variation at each school while leaving both stable between-school and stable between-course within-school differences in honors enrollment shares. It is motivated by the idea that school and department administrators may have idiosyncratic preferences or beliefs about honors efficacy that systematically shape their default choices of honors shares across years. Since there is a large persistent component of honors shares across years, the first stage is quite strong.[17] This specification yields point estimates that are noisier but also slightly (around .01SD) larger in magnitude than the baseline specification. The larger magnitudes could simply reflect either sampling error or a slight upward bias in the between-school variation that accounts for a larger share of identifying variation, but another possible explanation is that honors shares may be reported with error that is corrected by the IV specification, suggesting that the estimates from the baseline specification may be attenuated.

The third alternative specification alters the IV approach by using the mean honors fraction across all other courses in the same school-year as the instrument for the honors fraction in the chosen course. This specification removes systematic between-course variation in honors shares, leaving only persistent differences in schools' tendencies to have larger or smaller honors tracks as well as transitory cohort-specific variation in mean honors shares. Figure 4b shows that this second IV approach generally yields the same shapes and nearly

[17]The F statistics for the instruments for the first (linear) term in the cubic are all above 390, while their counterparts for the second (quadratic) and third (cubic) terms are all above 290 and 150, respectively. Estimates for the IV specification are produced using the "cmp" Stata function (Roodman, 2007).

the same magnitudes for the treatment effect functions as the baseline specification, albeit with slightly noisier point estimates, demonstrating that the general pattern of results does not hinge exclusively on the exogeneity of the residual between-course variation (about 47.4% of total residual baseline variation). This specification does feature a peak for quintile 3 at a slightly higher honors share, around 40%, than other specifications, which is consistent with the fact that a disproportionate number of quintile 3 students are near the honors margin around 40% relative to other quintiles.

Perhaps most importantly, though, the baseline specification and all three alternative specifications share four qualitative features: 1) students in the top quintiles benefit significantly from honors programs containing fewer than 30% of the student body; 2) students in the 2nd and 3rd quintiles benefit most from honors programs with 20-40% of the student body in them; 3) Students in the 4th quintile are relatively unaffected by changing the fraction of students in honors, with potentially small gains from small honors programs; and 4) students in quintile 5 are on average unaffected by honors programs with less than 40% of the student body in them and hurt by honors programs with more than 40% of the student body in them. As emphasized above, such consistency is unlikely to occur if endogeneity were driving the results, since different sources of endogeneity would need to cause the same pattern of bias across the interval of honors shares for all five quintiles of the preparedness distribution.

Interestingly, our results show that honors tracking programs are not zero sum. Small honors programs (between 25% and 40%) provide a Pareto improvement across quintiles relative to large honors programs ($> 40\%$), with some quintiles exhibiting sizable gains. Note that the decline in efficiency as honors tracks expand beyond 40% seems unlikely to be attributable to a consistent negative correlation with unobserved student quality that occurs at all three levels of variation. After all, recall from Table 3 that, if anything, the observed student characteristics (in particular the distribution of past test scores and parents' education) seem more favorable for school-year-course combinations featuring honors shares above .35. Thus, to generate the estimated decline spuriously, one would need mean values of observed and unobserved favorable student characteristics to be negatively correlated across school-year-courses featuring different honors shares, which would conflict with the predictions of standard models of student sorting.

One can potentially reconcile our results with papers finding that introducing tracking does not harm any students if the samples in those papers primarily contain schools that have small honors programs (Zimmer, 2003; Figlio and Page, 2002; Pischke and Manning, 2006). Similarly, one can also potentially reconcile our results with papers finding that honors programs help top students and hurt bottom students if those papers sampled a greater share

of schools with larger honors programs (Betts and Shkolnik, 2000; Hoffer, 1992; Argys et al., 1996; Epple et al., 2002).

5.2 Administrator's Problem

Armed with the estimates just presented of the quintile specific treatment effect functions $\{\hat{E}[\Delta\overline{Y}_q(f)]\}$, we can now reconsider the administrator's problem (2) from Section 2.1. Recall that solving for the optimal choice of honors selectivity also requires supplying weights $\{\theta_q\}$ capturing the relative importance the administrator places on achievement gains from each quintile of the student preparedness distribution. We consider two sets of weights. The first set weighs all quintiles equally ($\theta_q = \frac{1}{5}\ \forall q$), while the second set strongly prioritizes bottom quintiles, so that test score gains for quintiles 1, 2, 3, and 4 are weighted at 20%, 40%, 60%, and 80% of gains for quintile 5 respectively ($\theta_q = \frac{q}{15}\ \forall q$).[18]

The left panel of Figure 6 shows the average net student gains as a function of the honors fraction under equal weighting of quintiles, based on the estimates from the baseline specification.[19] The maximized gain of 0.04 SDs relative to the absence of tracking occur when honors tracks contain just over 20% and 30% of students. The right panel of Figure 6 displays weighted average gains with the second set of weights that prioritize students in bottom quintiles. Notably, the maximum weighted average gain still occurs at honors programs with enrollment shares between 20% and 30%, with a weighted average impact of 0.03 SDs. More generally, tracking schemes in which honors accounts for 20% to 30% of enrollment dominate those with larger honors tracks for any weighting scheme that places at least 10% of the weight on each of the five quintiles. In other words, further increases in the share of students in honors beyond 35% generate consistent decreases aggregate achievement gains for every reasonable weighting scheme over the remaining support of the data. The remarkable robustness of the optimal honors program size across weighting schemes is driven by gains for the top 60% of students from small honors programs and the lack of negative effect of small honors programs on students in the bottom 40% of the preparedness distribution.

The optimal size for an honors track is also robust across specifications. Figure 5 displays the average effect for the three alternate specifications under both weighting schemes. The school fixed effect specification, on the left side of Figure 5, has a smaller maximized weighted average gain, but the optimal share in honors remains around 20%. The lagged course-specific IV specification presented in Figure 5 has larger point estimates for the weighted

[18] Additional weighting schemes are available upon request from the authors.

[19] Confidence intervals for the values of the administrator's objective function are also generated using the delta method.

average gain, but the same optimal share of honors. The IV specification based on average honors shares from other courses from Figure 4b mimics the baseline specification closely, with the same weighted average gains of .04 SDs and .03 SDs under equal and compensatory weighting, respectively.

A .04 SD aggregate gain from introducing an honors track serving around 25% of students may seem relatively small; holding the baseline test score distribution fixed as a reference point, it would move a student at the statewide median to the 51.6^{th} percentile. However, this mean gain includes all students in the cohort for every course in which tracking is introduced. Also, the small value may be misleading given that the lion's share of achievement variance is determined by parents, innate student ability, and previous schools and teachers. Thus, it represents a considerable change in the value added of the high school. For example, using the estimates of Branch et al. (2012), it is equivalent to replacing a principal of median quality with one at the 64% of the principal quality distribution.

Furthermore, recent papers by Chetty et al. (2014a,b) and Carrell et al. (2018) analyzing changes in teacher quality and peer quality, respectively, have shown that policies generating modest short-run academic gains can produce substantial impacts on later life outcomes. Since teacher reallocation and specialization and changes in peer composition are two of the mechanisms through which honors track size is hypothesized to affect test scores, it seems plausible that tracking-induced achievement gains might similarly translate to later outcomes. While our data do not contain long-run outcomes of interest, we can perform a rough projection of the effect of our estimated test score gains on future earnings by assuming that test score gains from varying the size of honors programs have the same effect on age 28 earnings as the test score gains from improvements in teacher quality found in Chetty et al. (2014b,a).

Under this assumption, an initially trackless school that introduces optimally sized honors tracks for each core course in the sample could expect their students' earnings at age 28 to increase by an average of 0.4%.[20] For a high school class of 100 students near the age 28 median income, this implies an increase in aggregate age 28 earnings of over $88,000. This estimate would grow further if other courses not tested, such as English classes beyond English 1, enjoyed similar gains from tracking.

Of course, many schools already feature tracks near the optimal size for most of their courses. However, there remain a substantial share of school-year-courses in our sample that either do not use tracking or feature honors track sizes well outside the optimal range.

[20]This calculation assumes for simplicity that all students would have the 2018 median income at age 28 of $36,910 in the absence of tracking, and that test score gains from each subject can be translated to earnings gains and then aggregated across subjects.

If all schools in our sample switched from their current honors program size to an honors program with 20 to 30% of the student body in it, our estimates suggest that the average North Carolina student would experience a test score gain of over 0.02 SDs (about the same amount as switching from the median teacher to a 55th percentile teacher (Mansfield, 2015)). Since North Carolina averages about 100,000 students per cohort, this corresponds to an aggregate statewide increase in age 28 earnings of over $44 million.

Clearly, such back-of-the-envelope calculations are quite speculative; for example, they ignore general equilibrium effects from aggregate shifts in quality-adjusted labor supply as well as the substantial costs (and possible class size benefits) associated with staffing multiple tracks at small schools.[21] Nonetheless, they serve to highlight the possibility that small per-student gains from a superior tracking system can aggregate to very large earnings contributions when combining effects across many courses, schools, states, and years.

Limited or lack of benefit for the bottom quintile students could be addressed by reallocating resources to those students. These resources could include reduced class size for the regular track or a more targeted allocation of high-quality teachers to the regular track.

6 Robustness Checks

In order to maximize power, all of the results presented to this point have imposed that each quintile's expected achievement follows a restricted cubic function of the fraction of students in the honors track that takes on zero values at both ends of the unit interval. However, to demonstrate that our main findings are not driven primarily by assumptions about functional form, here we present results from several alternative specifications for the shape of $E[\Delta\overline{Y}_q(f_{stj})]$. Predicted treatment effects at candidate honors shares for all specifications are displayed in Table 5, while coefficient estimates are displayed in Table 2.

Figure A.6 plots a flexible semi-parametric specification that replaces the cubic specification with a set of interactions between student preparedness quintiles and quintiles of the fraction of students in honors:

$$E[\Delta\overline{Y}_q(f_{stj})] = \sum_{q'}\sum_{f'} 1(q = q')1(f_{stj} = f')\lambda_{q'f'} \tag{12}$$

Due to considerably greater imprecision, Figure A.6 plots estimated treatment effects along with 90% rather than 95% pointwise confidence intervals. Nonetheless, one can clearly see the same qualitative patterns for each quintile as Figure 3a. Specifically, for quintiles 1-4, expected gains compared to no tracking are generally above zero for honors shares between

[21] Note that a full welfare analysis also requires incorporating the effort costs paid by students. See Fu and Mehta (2018) for an example of a complete welfare assessment.

0% and 20%, then rise further between 20-40% before falling again for larger shares. The estimates for quintile 5 exhibit the same shape, but with negligible gains for small honors tracks relative to no tracking and meaningful losses when honors shares are so high that most students in this quintile are effectively in a remedial class.

Next, we consider relaxing the restriction that 100% of students in honors is equivalent to 0%. This addresses the possibility that a designation of "honors" connotes higher standards and a slightly more rigorous curriculum even when the student population is the same. Appendix Figure A.7 displays the results from an unrestricted cubic specification that fits three parameters per quintile. The shapes of the conditional expectation functions are quite similar over the range between 0% and 70% honors that spans nearly the entire support of the data, so that the specifications only meaningfully differ in their extrapolations to rarely-observed honors shares above 70%.

We also consider a specification that introduces a discontinuity at 0 to distinguish the absence of tracking from a very small tracking program:

$$E[\Delta\overline{Y}_q(f_{stj})] = \gamma_q^{lin} f_{stj} + \gamma_q^{sq} f_{stj}^2 - (\gamma_q^{lin} + \gamma_q^{sq}) f_{stj}^3 + \gamma_q^{indicator} \mathbb{1}_{(f_{stj} \in (0,1))} \tag{13}$$

Theoretically, this captures the possibility that teacher allocation and curriculum preparation may change discretely when even a tiny honors track exists. More practically, it ensures that the fitted values for smaller honors track sizes are not primarily being driven by the performance of students in untracked courses combined with a functional form that requires smoothness at 0. Appendix Figure A.8 shows that none of the quintiles features a discontinuity that is statistically or practically significant.

Appendix Figure A.9 considers a constrained quartic specification:

$$E[\Delta\overline{Y}_q(f_{stj})] = \gamma_q^{lin} f_{stj} + \gamma_q^{sq} f_{stj}^2 + \gamma_q^{cb} f_{stj}^3 - (\gamma_q^{lin} + \gamma_q^{sq} + \gamma_q^{cb}) f_{stj}^4 \tag{14}$$

The estimated predicted values are somewhat noisier than their cubic counterparts but are otherwise essentially unchanged.

We also consider two additional specifications that exchange reduced precision for arguably superior isolation of exogenous variation. First, inspired by the "Maimonides rule" identification strategy of Angrist and Lavy (1999) and others, we employ a specification that uses the share of classrooms that are assigned to the honors track as an instrument for the share of all students who take the honors track. Essentially, the high per-pupil staffing cost of offering class times with very few students may limit the set of viable honors fractions a school can choose. For example, a school with around 75 students in a cohort may have too many students for two classes and two few for four classes, so that the only feasible shares of honors classes are 0, .33, and .66. A larger cohort of 90 students might force the

school to allocate four classes, leading to honors class shares of 0, .25, .5, or .75. Thus, the discreteness inherent in forming classes may cause relatively small differences in cohort sizes to cause substantial arguably exogenous differences in honors shares (conditional on controls for class size). Appendix Figure A.10 displays the treatment effect functions from this additional IV specification. Again, the basic patterns remain the same for all quintiles.

Second, we augment our baseline specification with a full set of school-year combination fixed effects, so that estimates are identified exclusively by comparisons in relative performance across courses featuring different honors share within cohorts. These fixed effects are likely to remove almost all bias caused by student sorting, since these courses are being populated by nearly the same set of students.[22] Thus, any remaining bias would require either that the relative honors share responds to particular cohorts' unobserved mean comparative advantage in some subject (likely to be negligible) or that it responds to differential unobserved mean teacher quality or track-specific experience across courses (Cook and Mansfield, 2016). Appendix Figure A.11 displays results from this specification. The patterns are the same, but the effect sizes are somewhat muted, and the estimates are imprecise. One possible explanation for smaller impacts are that across-course differences in honors shares are likely to be quite small and transitory, and may not engender some of the teacher re-optimization of pace and pedagogical approach that drives gains from tracking.

Finally, it is possible that the school-year-courses chosen in section 3.1 are not sufficiently similar in their joint distributions of student abilities and costs to satisfy Assumption 1 and thus make the peer environment comparable in different schools featuring the same honors fraction. Thus, we re-estimate our baseline specification on a smaller subset of schools in which students' prior achievement would need to change by less than a third of a quintile on average to match the statewide uniform distribution of quintiles. Figure A.12 shows that the point estimates are nearly the same with the restricted sample, but with larger confidence intervals.

7 Conclusion

In this paper we use rich administrative data to identify the treatment effects of changing the size of the honors track, operationalized via functions of the share of students who enroll in honors, with separate functions estimated for each quintile of an index of student predicted performance. Importantly, our approach explicitly accommodates endogenous self-sorting of students into the honors and regular tracks conditional on the administrator-determined

[22] Some core courses, such as English 1 and Biology, are taken nearly universally, while others, such as Chemistry, are not taken by a substantial share of students.

capacity of the honors track. We then show that our set of estimated treatment effect functions suffice to determine the optimal share of students in each track in an administrator's planning problem.

We obtain the result that the optimal share of students in the honors track is between 20% and 30%. Based on results from our baseline specification, if all North Carolina public high schools switched from their current honors track sizes (including the absence of an honors program) to one with 20% to 30% of students in it, their students would on average gain over 0.02 SDs in test score performance compared to the existing statewide test score distribution. Altering the size of the honors track thus represents a low cost method to improve test score performance, particularly for larger schools that are already offering the relevant courses in several class periods. Importantly, the tradeoff between efficiency and equity is minimal, since highly prepared and moderately prepared students benefit considerably from small honors tracks (between 0.04 and 0.08 SDs) relative to the absence of tracking, while less prepared students only begin to experience losses when the honors track expands to nearly half the student population. Because these small per-student gains apply to such a wide population of students and high schools, our back-of-the-envelope calculations suggest that they could translate to aggregate skill development worth tens of millions of dollars in future earnings potential. To provide reassurance about the validity of these findings, we show that they are extremely robust across several alternative specifications featuring different samples, functional form assumptions, or segments of variation that remove different sources of endogeneity.

A few caveats about external validity are necessary. First, our approach assumes that students and their parents ultimately make track choices for each class, but that school administrators can alter incentives as necessary to induce their desired aggregate shares of students in each track. Thus, our results may not be externally valid for high schools where principals relinquish any role in shaping the honors track or for high schools where students can be assigned to tracks without their permission.

Similarly, our approach also requires drawing comparisons among the considerable majority of schools whose student populations feature distributions of past performance that minimally deviate from the statewide distribution. Thus, our results may not be externally valid to high schools with particularly large shares of very advanced or struggling students, since the peer composition in their honors or regular tracks may not be well-approximated by those at other schools, even conditional on the same student share in the honors track.

In addition, the North Carolina context we consider provides strong incentives to keep the breadth of material covered by the course similar among both tracks in order to prepare all students for a common statewide standardized exam. This feature is essential for

generating internally valid estimates by facilitating comparisons on a single achievement metric. However, we cannot verify external validity for contexts in which different tracks have substantially different curricula (e.g. Advanced Placement or International Baccalaureate), though we have no *a priori* reason to believe that our results would not generalize.

Finally, while our results may provide parents with a basis for comparing the tracking policies of schools they are considering, they are not intended to provide parents with information on whether or not their child should enroll in honors in a given course. This would require estimates of a different set of parameters that capture student-level treatment effects from switching tracks. A full decomposition of the effect from expanding the honors track into effects on the marginal students and peer effects in both the expanding and contracting tracks necessitates combining our estimates with exogenous variation in student-level track choices.

References

T. Ahn and J. Vigdor. The impact of no child left behind's accountability sanctions on school performance: Regression discontinuity evidence from north carolina. Technical report, National Bureau of Economic Research, 2014.

J. G. Altonji and R. K. Mansfield. Estimating group effects using averages of observables to control for sorting on unobservables: School and neighborhood effects. *American Economic Review*, 108(10):2902–46, 2018.

J. D. Angrist and V. Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly journal of economics*, 114(2):533–575, 1999.

D. Archbald and J. Keleher. Measuring conditions and consequences of tracking in the high school curriculum. *American Secondary Education*, pages 26–42, 2008.

L. M. Argys, D. I. Rees, and D. J. Brewer. Detracking america's schools: Equity at zero cost? *Journal of Policy analysis and Management*, pages 623–645, 1996.

J. R. Betts 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, 2000.

G. F. Branch, E. A. Hanushek, and S. G. Rivkin. Estimating the effect of leaders on public sector productivity: The case of school principals. Technical report, National Bureau of Economic Research, 2012.

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

S. E. Carrell, M. Hoekstra, and E. Kuka. The long-run effects of disruptive peers. *American Economic Review*, 108(11):3377–3415, 2018.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632, September 2014a. doi: 10.1257/aer.104.9.2593. URL http://www.aeaweb.org/articles?id=10.1257/aer.104.9.2593.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9):2633–79, 2014b.

J. B. Cook and R. K. Mansfield. Task-specific experience and task-specific talent: Decomposing the productivity of high school teachers. *Journal of Public Economics*, 140:51–72, 2016.

E. Duflo, 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(5):1739–74, 2011.

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

D. N. Figlio 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.

J. C. Fruehwirth. Identifying peer achievement spillovers: Implications for desegregation and the achievement gap. *Quantitative Economics*, 4(1):85–124, 2013.

C. Fu and N. Mehta. Ability tracking, school and parental effort, and student achievement: A structural model and estimation. *Journal of Labor Economics*, 36(4):923–979, 2018.

E. A. Hanushek et al. Does educational tracking affect performance and inequality? differences-in-differences evidence across countries. *The Economic Journal*, 116(510), 2006.

T. B. Hoffer. Middle school ability grouping and student achievement in science and mathematics. *Educational evaluation and policy analysis*, 14(3):205–227, 1992.

S. A. Imberman, A. D. Kugler, and B. I. Sacerdote. Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. *American Economic Review*, 102(5):2048–82, 2012.

L. Lefgren. Educational peer effects and the chicago public schools. *Journal of Urban Economics*, 56(2): 169–191, 2004.

M. C. Long, D. Conger, and P. Iatarola. Effects of high school course-taking on secondary and postsecondary success. *American Educational Research Journal*, 49(2):285–322, 2012.

R. K. Mansfield. Teacher quality and student inequality. *Journal of Labor Economics*, 33(3):751–788, 2015.

N. Mehta, R. Stinebrickner, and T. Stinebrickner. Time-use and academic peer effects in college. *Economic Inquiry*, 57(1):162–171, 2019. doi: 10.1111/ecin.12730. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12730.

J.-S. Pischke and A. Manning. Comprehensive versus selective schooling in england in wales: What do we know? Working Paper 12176, National Bureau of Economic Research, April 2006. URL http://www.nber.org/papers/w12176.

D. Roodman. CMP: Stata module to implement conditional (recursive) mixed process estimator. Technical report, Oct. 2007. URL https://ideas.repec.org/c/boc/bocode/s456882.html.

J. A. Smith and P. E. Todd. Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review*, 91(2):112–118, 2001.

G. Solon, S. J. Haider, and J. M. Wooldridge. What are we weighting for? *Journal of Human resources*, 50 (2):301–316, 2015.

J. L. Vigdor et al. Teacher salary bonuses in north carolina. In *Conference paper, National Center on Performance Incentives.-0.026*, 2008.

R. Zimmer. A new twist in the educational tracking debate. *Economics of Education Review*, 22(3):307–315, 2003.

J. D. Zinth. End-of-course exams. *Education Commission of the States (NJ3)*, 2012.

Tables

Table 1: Decomposing the Total and Residual Variance in Honors Enrollment Share

Variance Component	% of Total Variance $Var(f_{stj})$	% of Residual Variance $Var(f_{stj} - \overline{X}_{stj}\beta)$
Between School	37.7%	39.2%
Within School/Across Year	12.6%	13.5%
Within School-Year/Across Course	49.8%	47.4%

Notes: The subscripts s, t, and j denote school, year, and course, respectively. "Between School" captures $Var(\overline{f}_s)$ in Column 1 and $Var(\overline{f}_s - \overline{X}_s\beta)$ in Column 2. "Within School/Across Year" captures $Var(\overline{f}_{st}) - Var(\overline{f}_s)$ and $Var(\overline{f}_{st} - \overline{X}_{st}\beta) - Var(\overline{f}_s - \overline{X}_s\beta)$, respectively. "Within School-Year/Across Course" captures $Var(f_{stj}) - Var(\overline{f}_{st})$ and $Var(f_{stj} - \overline{X}_{stj}\beta) - Var(\overline{f}_{st} - \overline{X}_{st}\beta)$, respectively.

Table 2: Frequency of Tracking Offerings by Course

Course Name	No tracking	Only honors	Only remedial	Honors & remedial	Honors & AP	Only AP	Honors, AP, & remedial
Algebra 1	6872	588	131	16	0	0	0
Algebra 2	316	3695	3	9	0	0	0
Biology	422	4078	22	125	0	0	0
Chemistry	405	2230	0	0	0	0	0
English 1	179	4343	17	334	0	0	0
Geometry	466	3599	3	21	0	0	0
PSCI	2128	1190	102	83	0	0	0
Physics	30	416	0	0	129	18	0
US History	93	668	8	17	2149	245	47

Notes: Each cell provides the total number of school-year combinations in which the course indicated by the row title is offered under the tracking regime featured in the column titles. "PSCI" denotes physical science. The sample of school-years is limited to those with at least 30 test scores.

Table 3: Summary Statistics for Control Variables in X_{sjct} by Honors Enrollment Share

VARIABLES	No honors tracking mean (sd)	Share $\in (0, 0.35)$ mean (sd)	Share $\in [0.35, 1)$ mean (sd)
Title 1 status	0.977	0.987	0.986
	(0.150)	(0.115)	(0.119)
Cohort size	171.0	280.8	320.0
	(134.3)	(174.1)	(279.3)
Average class size	17.07	18.00	17.65
	(4.377)	(3.864)	(10.61)
Share of seats in remedial classes	0.00481	0.00281	0.00127
	(0.0356)	(0.0177)	(0.00943)
7th grade math scores	0.220	0.296	0.478
	(0.471)	(0.350)	(0.340)
8th grade math scores	0.544	0.613	0.776
	(0.508)	(0.375)	(0.356)
7th grade reading scores	0.198	0.255	0.417
	(0.389)	(0.286)	(0.267)
8th grade reading scores	0.502	0.557	0.712
	(0.384)	(0.283)	(0.264)
Average Praxis scores	541.0	544.0	527.7
	(180.6)	(148.1)	(168.0)
Teacher share with			
Bachelor's	0.901	0.890	0.872
	(0.246)	(0.225)	(0.249)
Master's	0.265	0.245	0.274
	(0.371)	(0.316)	(0.338)
Advanced degree	0.00595	0.00909	0.00831
	(0.0597)	(0.0655)	(0.0666)
Doctorate	0.00777	0.00214	0.00381
	(0.0761)	(0.0329)	(0.0474)
Standard professional II licenses	0.899	0.907	0.901
	(0.251)	(0.202)	(0.221)
Standard professional I licenses	0.0558	0.0547	0.0624
	(0.194)	(0.156)	(0.176)
Provisional licenses	0.0201	0.0154	0.0127
	(0.117)	(0.0891)	(0.0827)
Temporary licenses	0.0302	0.0261	0.0272
	(0.135)	(0.108)	(0.116)
0 years exp	0.0622	0.0493	0.0700
	(0.203)	(0.166)	(0.201)
1 year exp	0.0326	0.0322	0.0341
	(0.142)	(0.119)	(0.131)
2 years exp	0.0383	0.0350	0.0352
	(0.163)	(0.129)	(0.133)
3-5 years exp	0.112	0.107	0.0977
	(0.264)	(0.220)	(0.224)
6-11 years exp	0.217	0.211	0.213
	(0.348)	(0.299)	(0.316)
12+ years exp	0.538	0.566	0.550
	(0.422)	(0.368)	(0.386)
. Fraction of students			
Whose parents lack a HS diploma/GED	0.0667	0.0616	0.0429
	(0.0491)	(0.0396)	(0.0322)
Whose parents have a HS diploma	0.254	0.240	0.182
	(0.0927)	(0.0826)	(0.0818)
Whose parents have some college	0.122	0.121	0.120
	(0.0588)	(0.0472)	(0.0478)
Whose parents attended trade or business school	0.0517	0.0467	0.0401
	(0.0624)	(0.0541)	(0.0368)
Whose parents attended community college	0.210	0.203	0.186
	(0.0735)	(0.0646)	(0.0763)
Whose parents have a 4-year degree	0.205	0.224	0.283
	(0.0946)	(0.0856)	(0.0942)
Whose parents have graduate degrees	0.0749	0.0817	0.132
	(0.0644)	(0.0578)	(0.0882)
With gifted status	0.106	0.135	0.170
	(0.150)	(0.127)	(0.160)
With learning disabilities	0.0405	0.0357	0.0275
	(0.0429)	(0.0308)	(0.0276)
That are Hispanic	0.0408	0.0437	0.0467
	(0.0412)	(0.0410)	(0.0417)
That are black	0.236	0.251	0.262
	(0.184)	(0.178)	(0.174)
That are white	0.696	0.675	0.661
	(0.198)	(0.195)	(0.190)
That are Asian	0.0149	0.0135	0.0193
	(0.0227)	(0.0191)	(0.0199)
School-course-year-quintiles	9,393	27,093	18,539

Notes: Each entry provides mean values (and standard deviations in parentheses) for the control variable listed in the row label among all school-year-course observations. The sample here matches the one used for our baseline specification, which is limited to school-years with at least 30 test score observations and which feature typical distributions of student quality (See Section 3.1).

Table 4: Estimates of the Values of the Quintile-Specific Treatment Effect Functions $E[\Delta\overline{Y}_q(f)]$ at Several Candidate Honors Enrollment Fractions for the Baseline and Alternative Specifications

Specification	Share in honors	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
OLS	.15	.0695	.0479	.0321	.0267	.0118
		(.0434, .0956)	(.0255, .0703)	(.0102, .0541)	(.0053, .0482)	(-.0079, .0316)
	.3	.0748	.0535	.0396	.0255	-.0005
		(.0424, .107)	(.0249, .082)	(.0113, .0678)	(-.0018, .0528)	(-.0256, .0245)
	.45	.0413	.0331	.0309	.0083	-.0242
		(.0074, .0753)	(.0020, .0642)	(.0004, .0613)	(-.0210, .0375)	(-.0513, .0029)
	.6	-.0053	.0030	.0147	-.0131	-.0464
		(-.0447, .0342)	(-.0332, .0393)	(-.0200, .0493)	(-.0466, .0203)	(-.0778, -.0149)
OLS School FEs	.15	.0448	.0253	.0141	.0088	-.0035
		(.0223, .0672)	(.0066, .0441)	(-.0047, .0329)	(-.0096, .0273)	(-.0225, .0155)
	.3	.0458	.0267	.0184	.0048	-.0169
		(.0170, .0746)	(.0015, .0519)	(-.0065, .0433)	(-.0199, .0295)	(-.0424, .0085)
	.45	.0210	.0138	.0159	-.0057	-.0332
		(-.0085, .0505)	(-.0134, .0409)	(-.0104, .0423)	(-.0323, .0208)	(-.0614, -.0049)
	.6	-.0116	-.0038	.0098	-.0167	-.0449
		(-.0432, .0200)	(-.0328, .0252)	(-.0181, .0378)	(-.0450, .0117)	(-.0766, -.0132)
Lagged IV	.15	.0943	.0651	.0539	.0434	.024
		(.0666, .1220)	(.0374, .0929)	(.0261, .0816)	(.0156, .0711)	(-.0037, .0518)
	.3	.1010	.0684	.0600	.0425	.0090
		(.0671, .1350)	(.0342, .1030)	(.0258, .0942)	(.0083, .0767)	(-.0251, .0432)
	.45	.0557	.0348	.0369	.0160	-.0253
		(.0193, .0921)	(-.0016, .0712)	(.0006, .0733)	(-.0204, .0524)	(-.0617, .0111)
	.6	-.0078	-.0106	.0030	-.0176	-.0594
		(-.0512, .0356)	(-.0540, .0328)	(-.0403, .0464)	(-.0610, .0258)	(-.1030, -.0160)
Other-course IV	.15	.0667	.0464	.0289	.0217	.0177
		(.0362, .0972)	(.0159, .0769)	(-.0016, .0595)	(-.0088, .0522)	(-.0129, .0482)
	.3	.0747	.0575	.0439	.0255	.0103
		(.0378, .1120)	(.0207, .0944)	(.0070, .0808)	(-.0114, .0623)	(-.0266, .0471)
	.45	.0465	.0455	.0472	.0179	-.0101
		(.0084, .0847)	(.0074, .0836)	(.0091, .0853)	(-.0202, .0561)	(-.0482, .0280)
	.6	.0049	.0225	.0414	.0057	-.0313
		(-.0406, .0504)	(-.0230, .0680)	(-.0040, .0869)	(-.0398, .0511)	(-.0768, .0142)

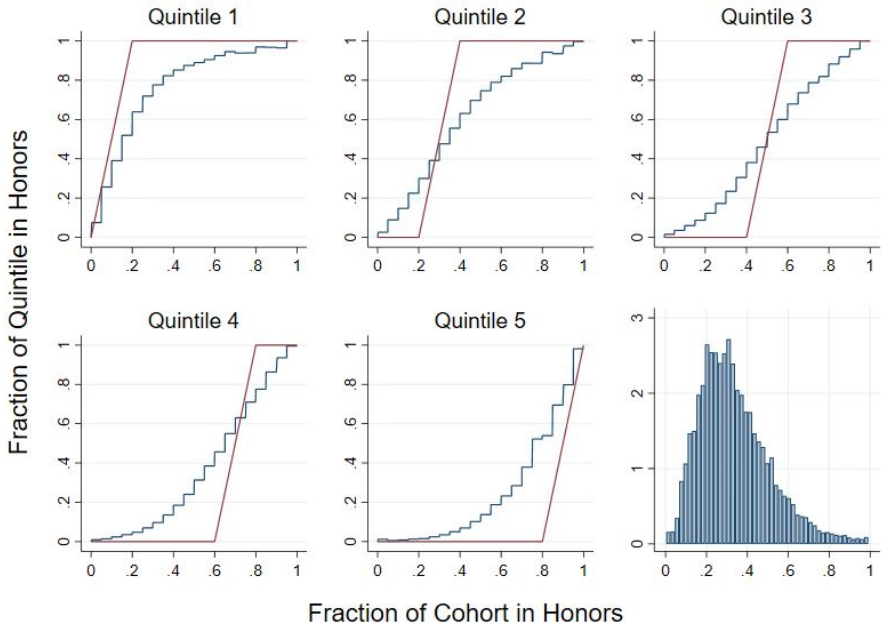
Notes: Predicted values are generated from the estimates $\hat{\vec{\gamma}}$ for the specifications in the row category for the values of the honors enrollment share f listed in the row labels. 95% confidence intervals computed using the delta method are displayed in parentheses. Each column presents estimates for a different quintile of the statewide predicted performance distribution among students. "OLS" refers to the baseline specification that pools all sources of variation in the honors enrollment share. "OLS School FEs" uses a full set of school fixed effects to isolate within-school variation. "Lagged IV" uses the previous year's honors enrollment share (and its square) as instruments for its contemporary counterparts in the chosen school-year-course. "Other-course IV" uses the contemporaneous honors enrollment share (and its square) in the other tested courses as instruments for the share and its square in the chosen course.

Table 5: Estimates of the Values of the Quintile-Specific Treatment Effect Functions $E[\Delta\overline{Y}_q(f)]$ at Several Candidate Honors Enrollment Fractions for Various Specifications Examining the Robustness of Results

Specification	Share in honors	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Bin Specification	[0, .2)	.0553	.0313	.0033	.0147	-.0006
		(.0146, .0960)	(-.0023, .0649)	(-.0298, .0365)	(-.0178, .0473)	(-.0311, .0299)
	[.2, .4)	.0686	.0470	.0187	.0126	-.0210
		(.0331, .1040)	(.0162, .0779)	(-.0106, .048)	(-.0160, .0412)	(-.0473, .0053)
	[.4, .6)	.0386	.0271	.0102	-.0072	-.0511
		(-.0011, .0784)	(-.0087, .0630)	(-.0234, .0439)	(-.0408, .0264)	(-.0828, -.0193)
	[.6, 1]	-.0135	-.0165	-.0144	-.0340	-.0685
		(-.0665, .0395)	(-.0645, .0316)	(-.0630, .0342)	(-.0830, .0151)	(-.1150, -.0215)
Unconstrained OLS	.15	.0615	.0425	.0310	.0291	.0144
		(.0324, .0906)	(.0177, .0673)	(.0073, .0547)	(.0058, .0524)	(-.0076, .0364)
	.3	.0662	.0476	.0383	.0280	.0021
		(.0304, .102)	(.0161, .0791)	(.0081, .0684)	(-.0016, .0575)	(-.0252, .0294)
	.45	.0347	.0283	.0297	.0101	-.0222
		(-.0002, .0714)	(-.0053, .0619)	(-.0024, .0619)	(-.0211, .0414)	(-.0509, .0064)
	.6	-.0126	-.0023	.0133	-.0110	-.0440
		(-.0548, .0296)	(-.0412, .0367)	(-.0238, .0504)	(-.0471, .0250)	(-.0775, -.0106)
Honors Indicator OLS	.15	.0592	.0367	.0266	.0277	.0144
		(.0205, .0979)	(.0033, .0701)	(-.0050, .0582)	(-.0034, .0589)	(-.0150, .0438)
	.3	.0703	.0486	.0371	.0258	.0004
		(.0341, .107)	(.0168, .0805)	(.0071, .0670)	(-.0039, .0556)	(-.0268, .0276)
	.45	.0373	.0283	.0283	.0085	-.0233
		(-.0006, .0752)	(-.0063, .0630)	(-.0041, .0607)	(-.0237, .0408)	(-.0528, .0062)
	.6	-.0127	-.0058	.0099	-.0125	-.0444
		(-.0603, .0349)	(-.0495, .0378)	(-.0307, .0504)	(-.0532, .0283)	(-.0818, -.0069)
Quartic	.15	.0672	.0423	.0275	.0196	.0044
		(.0324, .1020)	(.0124, .0722)	(-.0012, .0561)	(-.0083, .0474)	(-.0224, .0313)
	.3	.0739	.0518	.0381	.0235	-.0028
		(.0398, .1080)	(.0219, .0816)	(.0091, .0672)	(-.0047, .0516)	(-.0288, .0233)
	.45	.0412	.0331	.0307	.0082	-.0244
		(.0072, .0751)	(.0020, .0642)	(.0003, .0612)	(-.0211, .0375)	(-.0516, .0028)
	.6	-.0071	-.0017	.0105	-.0198	-.0531
		(-.0529, .0387)	(-.0435, .0402)	(-.0290, .0500)	(-.0586, .0191)	(-.0887, -.0175)
Class share IV	.15	.0690	.0465	.0320	.0226	.0107
		(.0473, .0907)	(.0248, .0682)	(.0103, .0537)	(.0009, .0443)	(-.0111, .0324)
	.3	.0722	.0517	.0377	.0206	-.0032
		(.0449, .0996)	(.0244, .0791)	(.0104, .0650)	(-.0067, .0480)	(-.0305, .0241)
	.45	.0363	.0317	.0267	.0047	-.0283
		(.0071, .0656)	(.0025, .0610)	(-.0026, .0559)	(-.0245, .0340)	(-.0575, .0010)
	.6	-.0121	.0024	.0086	-.0144	-.0512
		(-.0462, .0219)	(-.0317, .0364)	(-.0254, .0427)	(-.0485, .0196)	(-.0852, -.0171)
School-Year FEs	.15	.0388	.0207	.0081	.0021	-.0105
		(.0162, .0615)	(.0013, .0402)	(-.0122, .0285)	(-.0174, .0215)	(-.0302, .0093)
	.3	.0375	.0217	.0105	-.0041	-.0254
		(.008, .067)	(-.0047, .0481)	(-.0166, .0377)	(-.0303, .0222)	(-.0521, .0012)
	.45	.0130	.0109	.0091	-.0136	-.0397
		(-.0184, .0443)	(-.0182, .0401)	(-.0201, .0382)	(-.0424, .0152)	(-.0694, -.0100)
	.6	-.0177	-.0036	.0055	-.0217	-.0479
		(-.0520, .0165)	(-.0352, .0280)	(-.0258, .0368)	(-.0532, .0097)	(-.0809, -.0149)

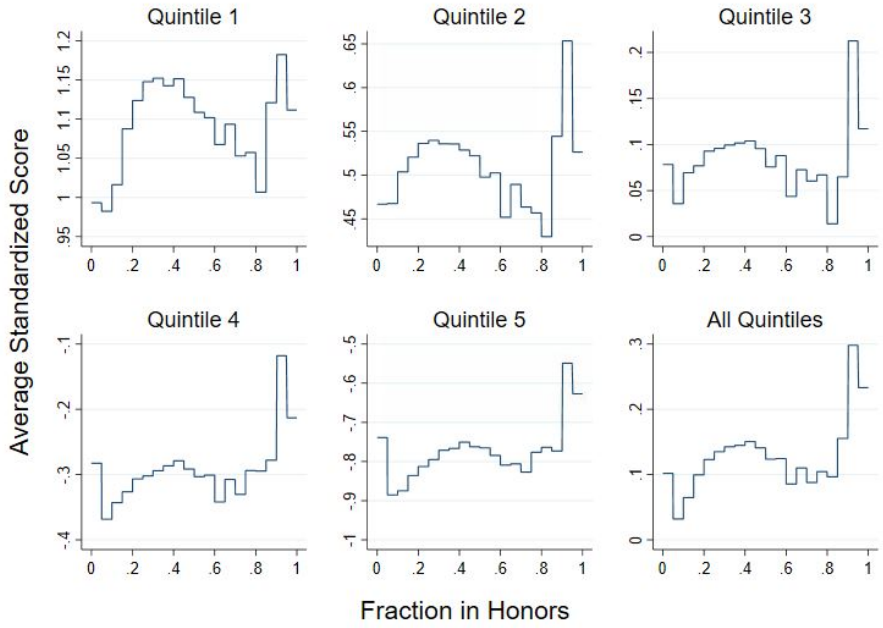
Notes: Predicted values are generated from the estimates $\hat{\tilde{\gamma}}$ for the specifications in the row category for the values of the honors enrollment share f listed in the row labels. 95% confidence intervals computed using the delta method are displayed in parentheses. Each column presents estimates for a different quintile of the statewide predicted performance distribution among students. "Bin Specification" alters the baseline specification by replacing the restricted cubic function with separate indicators for whether the share of students in honors falls within mutually exclusive intervals of length 0.2. "Unconstrained OLS" removes the restriction that the treatment effects for 0% and 100% honors enrollment are equal. "Honors Indicator OLS" alters the baseline specification by including a separate indicator for an honors enrollment share of 0. "Quartic" fits a quartic rather than a cubic polynomial while maintaining the restriction that the treatment effects for 0% and 100% honors enrollment are equal. "Class share IV" uses the honors classroom share (and its square) as instruments for the coursewide honors enrollment share (and its square). "School-Year FEs" introduces a full set of fixed effects for school-year combinations.

Figure 1: Student Probability of Choosing the Honors Track as a Function of the Coursewide Honors Enrollment Share by Quintile of the School-Specific Predicted Performance Distribution



Notes: The first five graphs plot the share of students in the chosen quintile of predicted performance that selects the honors track among narrow bins of the coursewide honors enrollment share. Quintiles for this figure are based on school-specific rather than statewide predicted performance rankings. Each bin includes shares in (bin minimum, bin maximum]. The bottom right cell plots the support of the data used for the other five cells, excluding school-year-courses where either none of the students or all of the students are enrolled in honors. The figures are based on the final sample of school-course-year-quintile observations used to estimate the baseline specification.

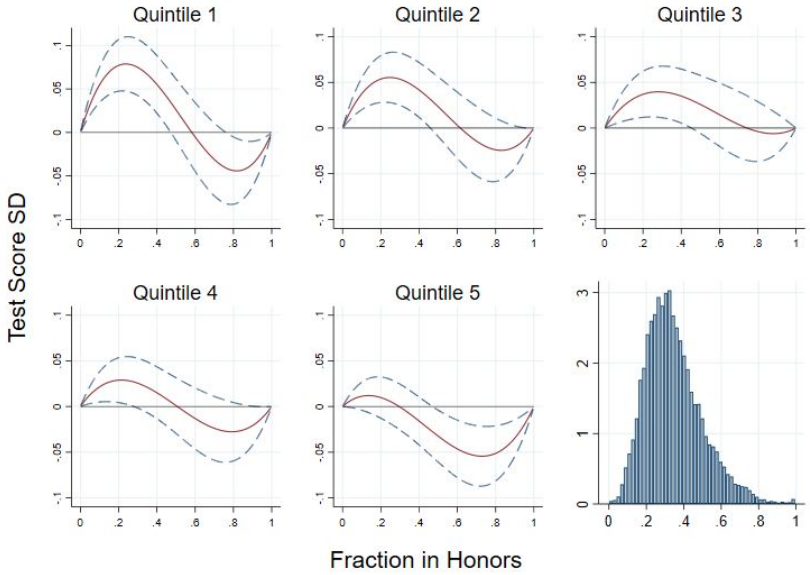
Figure 2: Average Standardized Score as a Function of the Coursewide Honors Enrollment Share by Quintile of Student Predicted Performance



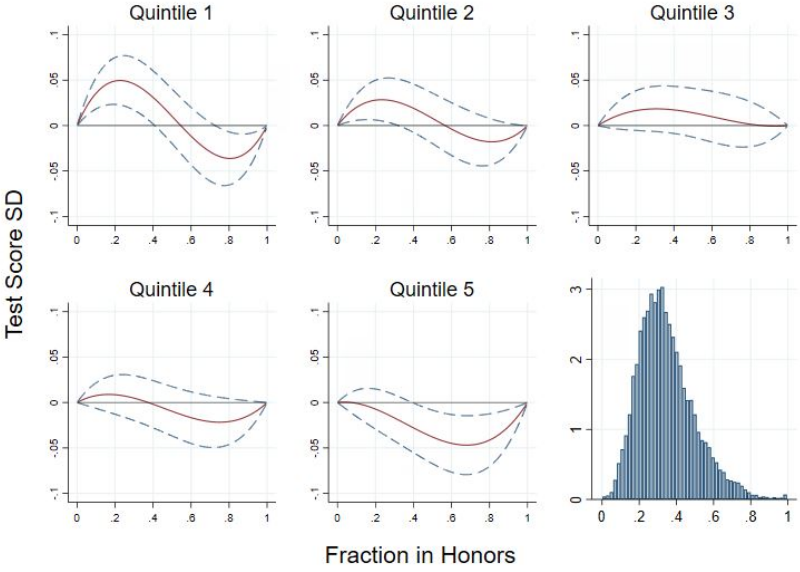
Notes: Each graph plots the mean standardized test score by narrow bins of the share of the course's students enrolled in honors (pooled across six subjects) for a different quintile of a regression index of predicted student performance based on grade 7 and 8 test scores. The bin for the lowest share of students in honors includes school-year-courses where no tracking occurs. The remaining bins consider honors enrollment shares in the interval (bin minimum, bin maximum].

Figure 3: Treatment Effect Functions for the Honors Enrollment Fraction by Quintile of Predicted Student Achievement ($E[\Delta\overline{Y}_q(f)]$): Baseline and School Fixed Effects Specifications

(a) Baseline



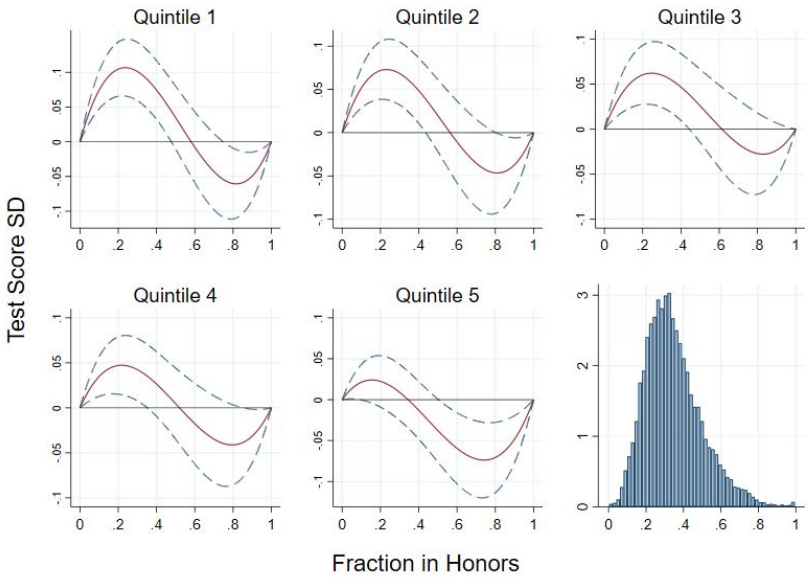
(b) School Fixed Effects



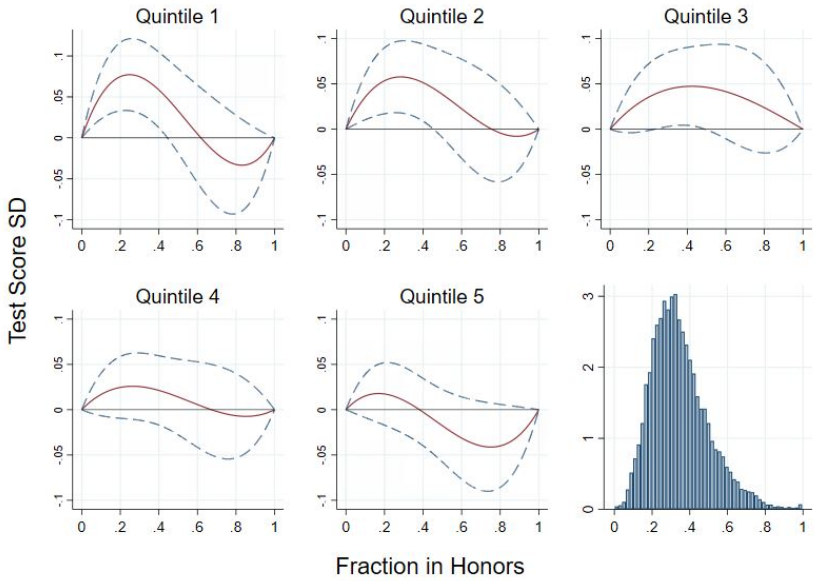
Notes: The first five graphs in each panel plot estimates of the function $E[\Delta\overline{Y}_q(f)]$ that maps coursewide honors enrollment fraction into expected standardized test performance by quintile of predicted performance for the baseline (panel (a)) or school fixed effects (panel (b)) specifications. The bottom right graph in each panel displays the density of honors enrollment shares for the baseline sample. The sample is restricted to school-year-course combinations that serve at least 30 students, do not offer IB nor AP tracks, and whose schools' distributions of student preparedness closely resemble the statewide distribution (See Section 3.1). 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes.

Figure 4: Treatment Effect Functions for the Honors Enrollment Fraction by Quintile of Predicted Student Achievement ($E[\Delta\overline{Y}_q(f)]$): IV Specifications

(a) IV (Previous Year's Honors Fraction)



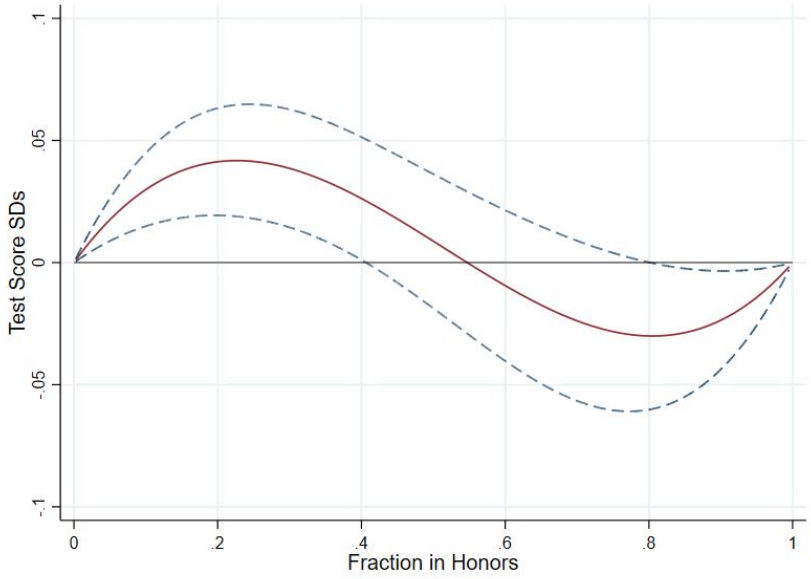
(b) IV (Other Courses' Mean Honors Fraction)



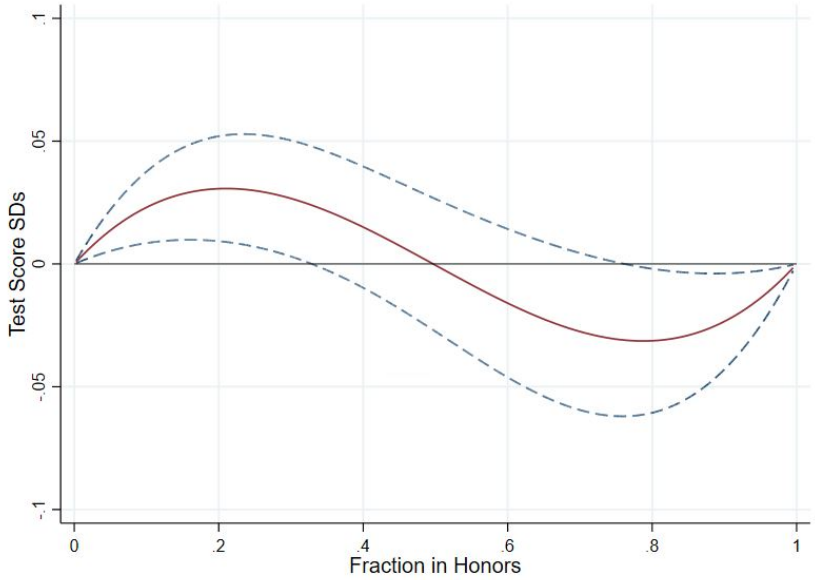
Notes: Panel (a) displays estimates of the function $E[\Delta\overline{Y}_q(f)]$ that maps coursewide honors enrollment fraction into expected standardized test performance by quintile of student predicted performance for a specification in which the current course's honors enrollment share is instrumented with the previous year's share. Panel (b) plots analogous estimates for a specification in which the current course's honors enrollment share is instrumented with the mean share among other courses in the same school-year combination. Both figures use the baseline sample of school-course-year-quintile observations (See Section 3.1 for details). The bottom right graph in each panel displays the sample's density of honors enrollment shares. 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes.

Figure 5: School Average Test Score Gains as a Function of the Honors Enrollment Fraction Using Equal vs. Compensatory Weights: Baseline Specification

(a) Equal Weighting



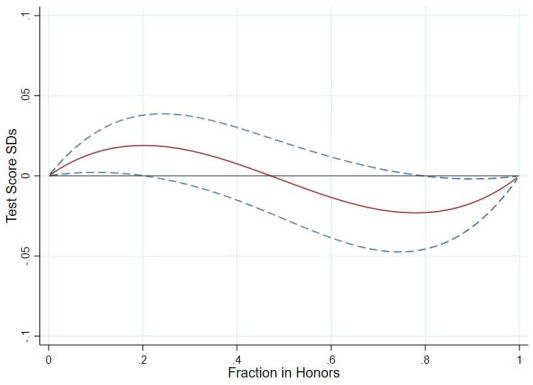
(b) Compensatory Weighting



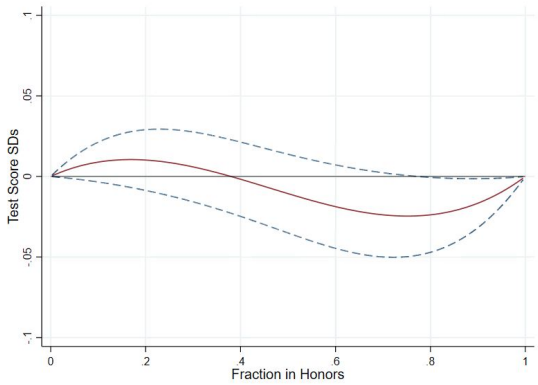
Notes: Each figure displays estimates of the value of the administrator's objective $\max_f \sum_{q=1}^{Q} W_q \theta_q E[\Delta \overline{Y}_q(f)]$ as a function of the coursewide honors enrollment fraction, where W_q is the share of the course's students who belong to the q-th predicted performance quintile and θ_q is the preference weight given to the achievement of quintile q. The left two graphs use estimates of $E[\Delta \overline{Y}_q(f)]$ from the baseline specification. "Equal Weighting": test scores gains by all quintiles are weighted equally. "Compensatory Weighting": quintiles 1, 2, 3, 4, and 5 are assigned weight $\frac{1}{15}$, $\frac{2}{15}$, $\frac{3}{15}$, $\frac{4}{15}$, and $\frac{5}{15}$. Each figure relies on the baseline sample of school-course-year-quintile observations (See Section 3.1 for details). 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes.

Figure 6: School Average Test Score Gains as a Function of the Honors Enrollment Fraction Using Equal vs. Compensatory Weights: Baseline Specification

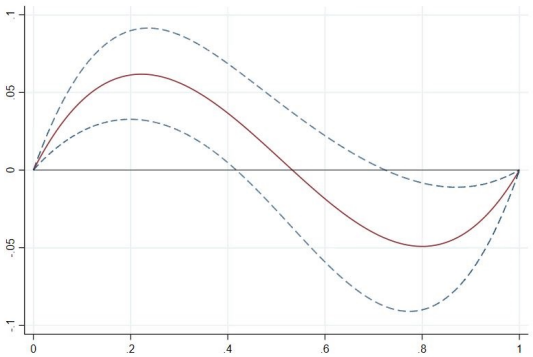
(a) School FEs: Equal Weighting



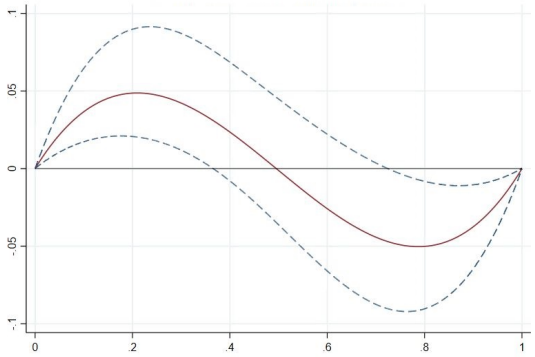
(b) School FEs: Compensatory Weighting



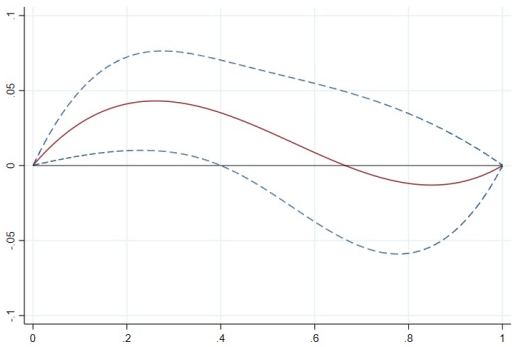
(c) Lagged IV: Equal Weighting



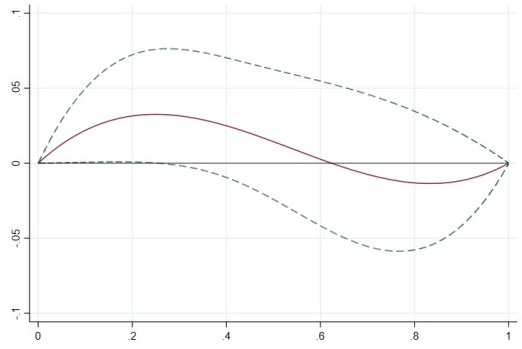
(d) Lagged IV: Compensatory Weighting



(e) Other Course IV: Equal Weighting



(f) Other Course IV: Compensatory Weighting



Notes: Each figure displays estimates of the value of the administrator's objective $\max_f \sum_{q=1}^{Q} W_q \theta_q E[\Delta \overline{Y}_q(f)]$ as a function of the coursewide honors enrollment fraction, using estimates of treatment effects $E[\Delta \overline{Y}_q(f)]$ from the alternative specification listed in the subtitle (See Section 4 for details). "Equal Weighting": test scores gains by all quintiles are weighted equally. "Compensatory Weighting": quintiles 1, 2, 3, 4, and 5 are assigned weight $\frac{1}{15}$, $\frac{2}{15}$, $\frac{3}{15}$, $\frac{4}{15}$, and $\frac{5}{15}$. Both figures use the baseline sample of school-course-year-quintile observations (See Section 3.1 for details). 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes.

A Appendix

Table 1: Estimates of the Parameters $\{\gamma\}$ Governing the Quintile-Specific Treatment Effect Functions of the Honors Enrollment Fraction $E[\Delta\overline{Y}_q(f)]$ for the Baseline and Alternative Specifications

	(1)	(2)	(3)	(4)
Quintile 1-Linear Coefficient	0.734***	0.484***	0.997***	0.691***
	(0.138)	(0.115)	(0.185)	(0.200)
Quintile 1-Squared Coefficient	-1.995***	-1.372***	-2.713***	-1.808***
	(0.401)	(0.314)	(0.542)	(0.607)
Quintile 1-Cubic Coefficient	1.260***	0.888***	1.716***	1.117***
	(0.276)	(0.210)	(0.373)	(0.424)
Quintile 2-Linear Coefficient	0.497***	0.270***	0.696***	0.454***
	(0.117)	(0.0940)	(0.154)	(0.169)
Quintile 2-Squared Coefficient	-1.303***	-0.747**	-1.929***	-1.055***
	(0.341)	(0.251)	(0.471)	(0.519)
Quintile 2-Cubic Coefficient	0.807**	0.477**	1.233***	0.601***
	(0.236)	(0.169)	(0.328)	(0.365)
Quintile 3-Linear Coefficient	0.316***	0.134	0.559***	0.245
	(0.114)	(0.0945)	(0.154)	(0.169)
Quintile 3-Squared Coefficient	-0.740**	-0.289	-1.470***	-0.366
	(0.325)	(0.252)	(0.455)	(0.492)
Quintile 3-Cubic Coefficient	0.424*	0.155	0.911***	0.121
	(0.224)	(0.169)	(0.316)	(0.341)
Quintile 4-Linear Coefficient	0.298***	0.115	0.478***	0.219
	(0.112)	(0.0925)	(0.150)	(0.160)
Quintile 4-Squared Coefficient	-0.886***	-0.423*	-1.396***	-0.544
	(0.321)	(0.248)	(0.448)	(0.474)
Quintile 4-Cubic Coefficient	0.588***	0.308*	0.919***	0.325
	(0.221)	(0.167)	(0.312)	(0.328)
Quintile 5-Linear Coefficient	0.188*	0.0259	0.334**	0.228
	(0.103)	(0.0964)	(0.148)	(0.164)
Quintile 5-Squared Coefficient	-0.824***	-0.381	-1.303***	-0.826*
	(0.299)	(0.269)	(0.438)	(0.484)
Quintile 5-Cubic Coefficient	0.636***	0.355*	0.969***	0.598*
	(0.207)	(0.186)	(0.304)	(0.334)
Observations	108,977	108,977	108,994	108,982
School FEs	NO	YES	NO	NO
Constrained Coefficients	YES	YES	YES	YES
Lagged IV	NO	NO	YES	NO
Other-course IV	NO	NO	NO	YES

Notes: *** p<0.01, ** p<0.05, * p<0.1. Robust standard errors clustered at the school level are in parentheses. "Lagged IV": instruments for the current course's honors share using the prior year's share. "Other-course IV": instruments for the current course's honors share using the mean contemporaneous share in the other courses.

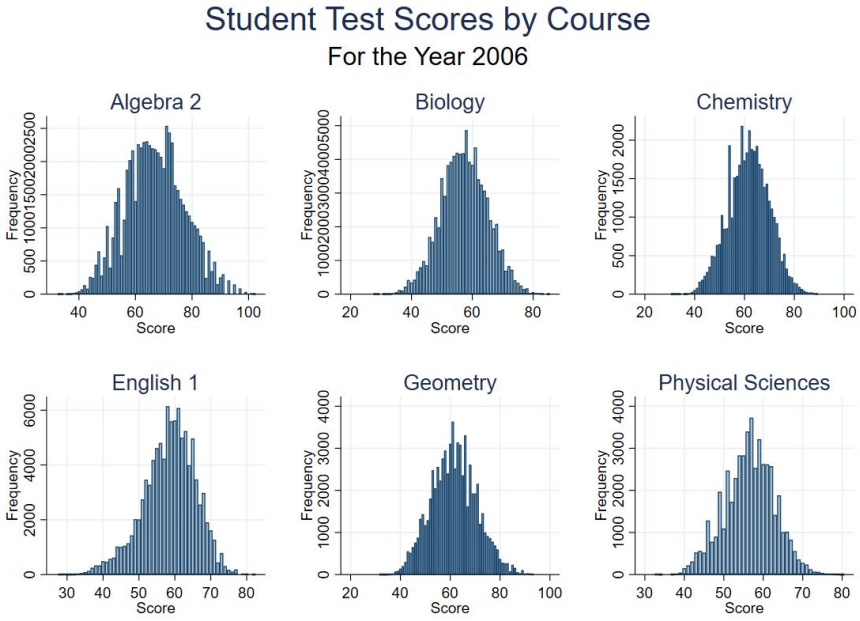
Table 2: Estimates of the Parameters $\{\gamma\}$ Governing the Quintile-Specific Treatment Effect Functions of the Honors Enrollment Fraction $E[\Delta\overline{Y}_q(f)]$ for Several Specifications Testing Robustness to Functional Form and Endogeneity Assumptions

	(1)	(2)	(3)	(4)
Quintile 1-Linear Coefficient	0.645***	0.847***	0.431***	0.738***
	(0.154)	(0.189)	(0.116)	(0.141)
Quintile 1-Squared Coefficient	-1.719***	-2.186***	-1.272***	-2.054***
	(0.442)	(0.470)	(0.318)	(0.414)
Quintile 1-Cubic Coefficient	1.014***	1.339***	0.841***	1.315***
	(0.317)	(0.298)	(0.216)	(0.287)
Quintile 2-Linear Coefficient	0.437***	0.625***	0.222**	0.483***
	(0.130)	(0.170)	(0.0968)	(0.123)
Quintile 2-Squared Coefficient	-1.121***	-1.529***	-0.617**	-1.271***
	(0.370)	(0.414)	(0.260)	(0.367)
Quintile 2-Cubic Coefficient	0.644**	0.904***	0.395**	0.788***
	(0.267)	(0.261)	(0.177)	(0.257)
Quintile 3-Linear Coefficient	0.303**	0.380**	0.0774	0.323***
	(0.124)	(0.186)	(0.102)	(0.123)
Quintile 3-Squared Coefficient	-0.704**	-0.856**	-0.168	-0.801**
	(0.357)	(0.424)	(0.275)	(0.359)
Quintile 3-Cubic Coefficient	0.392	0.476*	0.0906	0.478*
	(0.265)	(0.256)	(0.186)	(0.249)
Quintile 4-Linear Coefficient	0.324***	0.284*	0.0517	0.257**
	(0.122)	(0.170)	(0.0977)	(0.121)
Quintile 4-Squared Coefficient	-0.969***	-0.862**	-0.289	-0.784**
	(0.346)	(0.406)	(0.265)	(0.352)
Quintile 4-Cubic Coefficient	0.663***	0.578**	0.237	0.528**
	(0.248)	(0.251)	(0.181)	(0.243)
Quintile 5-Linear Coefficient	0.217*	0.155	-0.0428	0.183
	(0.117)	(0.162)	(0.0998)	(0.114)
Quintile 5-Squared Coefficient	-0.917***	-0.765**	-0.218	-0.842**
	(0.345)	(0.378)	(0.276)	(0.333)
Quintile 5-Cubic Coefficient	0.721***	0.610***	0.261	0.660***
	(0.259)	(0.233)	(0.190)	(0.230)
Observations	108,977	108,977	108,977	108,977
School FEs	NO	NO	NO	NO
School-Year FEs	NO	NO	YES	NO
Constrained Coefficents	NO	YES	YES	YES
Class Share IV	NO	NO	NO	YES
Honors Indicator	NO	YES	NO	NO
Sorting Metric<.5	YES	YES	YES	YES

Notes: *** p<0.01, ** p<0.05, * p<0.1. Robust standard errors clustered at the school level are in parentheses. "Class Share IV": instruments for the current course's honors enrollment share using its honors classroom share. "Honors Indicator": includes an indicator for the existence of tracking.

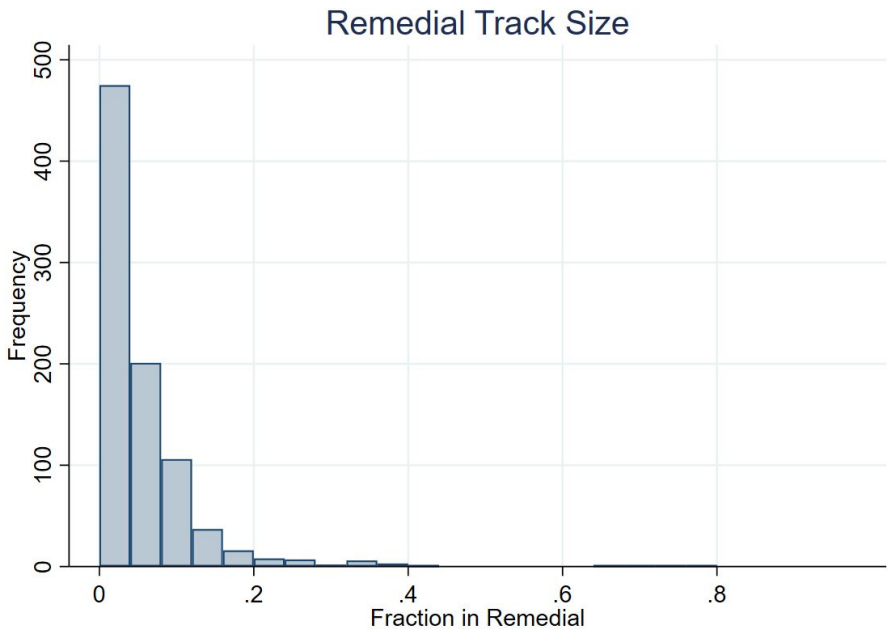
Figures

Figure A.1: Confirming the Absence of Floor and Ceiling Effects: The 2006 Empirical Distribution of Pre-Standardized Scale Scores for the Sample Courses



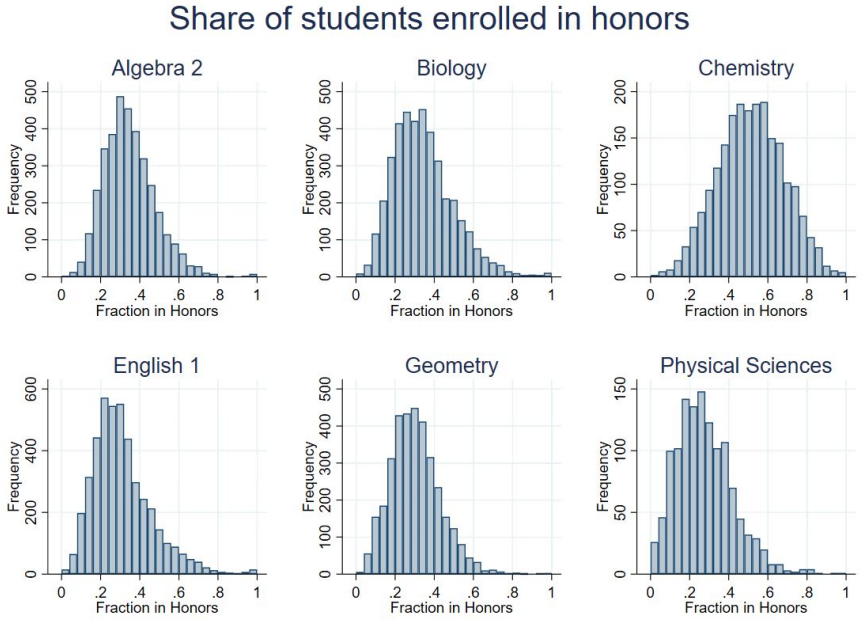
Notes: Each histogram depicts the distribution of pre-standardized student scale scores for the courses included in the final sample for the year 2006. The histograms confirm the absence of bunching near the ceiling or floor of the test score range. More years are available by request.

Figure A.2: The Distribution of Remedial Enrollment Shares



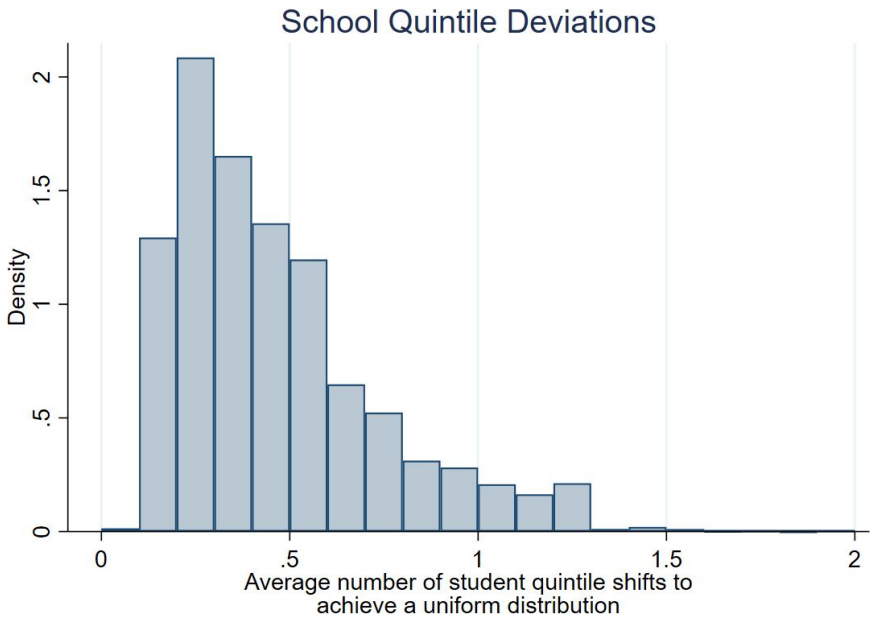
Notes: This figure depicts the fraction of students in the remedial track for school-year-courses from the baseline sample in which a remedial track exists. Fewer than 4% of school-year-courses in the sample have a remedial track.

Figure A.3: The Distribution of Honors Track Enrollment Shares by Course



Notes: Each figure depicts the distribution of the fraction of students who enroll in the honors track for school-year-courses in which an honors track exists for the labeled course. The figures rely on the baseline sample of school-course-year observations (See Section 3.1 for details).

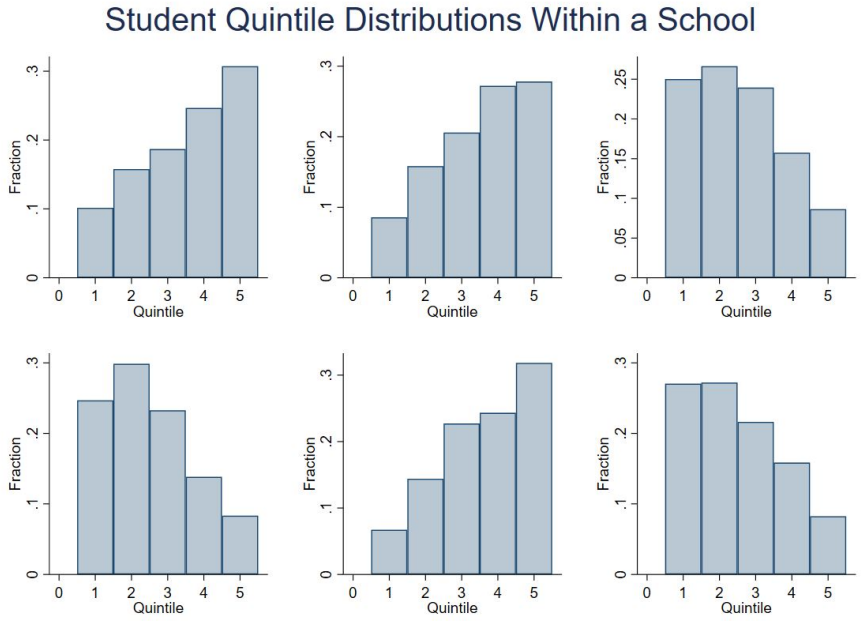
Figure A.4: Assessing the Validity of Assumption 1: The Distribution of School-Specific of Departures from the Statewide Composition of Student Predicted Performance



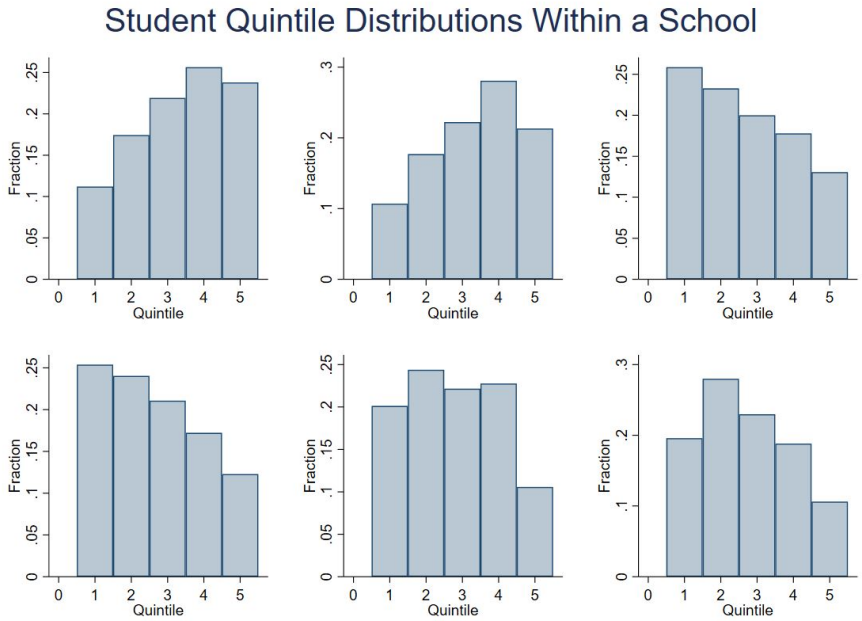
Notes: This figure displays the distribution among high schools of the average number of quintiles of an index of predicted test score performance by which the school's students would need to be shifted to match the statewide (uniform) distribution of student predicted performance quintiles. Larger values indicate that the school's student population is more atypical.

Figure A.5: The Distribution of Student Predicted Performance Quintiles for the Schools on the Margin of Sample Inclusion

(a) 0.5 Quintile Shifts/Student

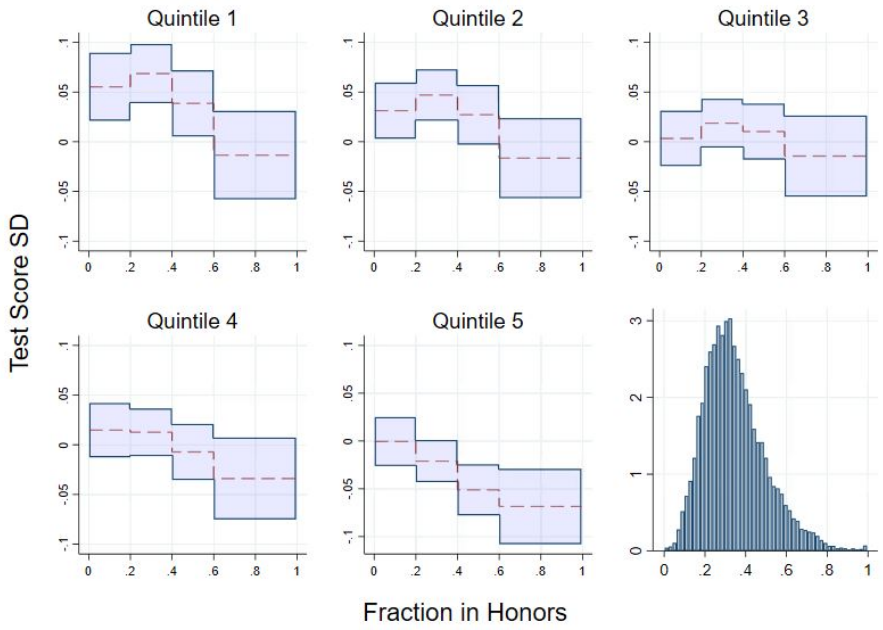


(b) 0.33 Quintile Shifts/Student



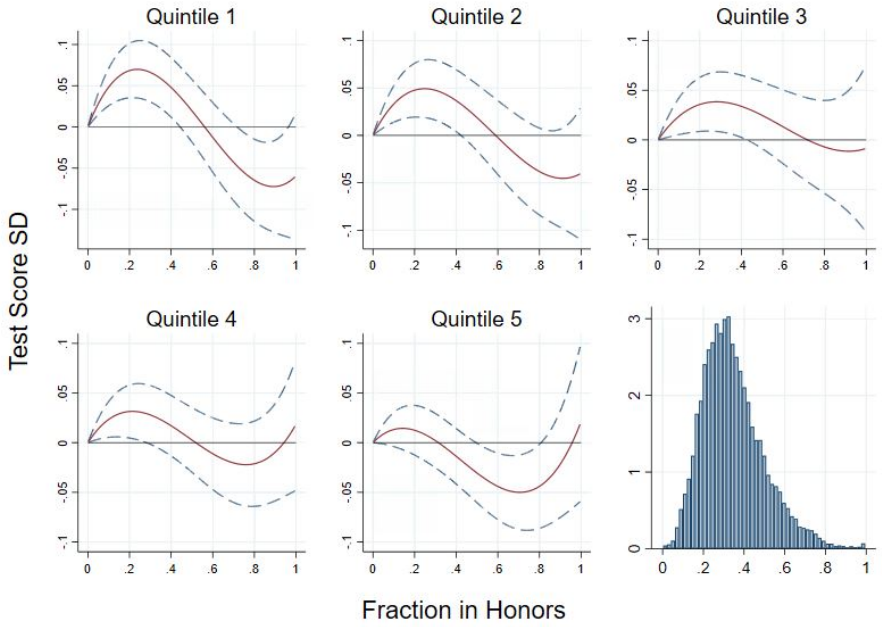
Notes: Figure (a) displays the distribution of students classified by statewide quintile of the regression index of predicted test scores for the six schools with the highest deviations from the statewide (uniform) distribution of quintiles that still qualified for the baseline sample (0.5 required quintile shifts per student on average to reach the uniform distribution). Figure (b) plots the distributions for the six marginal schools when the standard is lowered to one-third quintile shifts per student.

Figure A.6: Testing Robustness to Alternative Functional Forms for the Treatment Effect Functions $E[\Delta\overline{Y}_q(f)]$: Estimating Quintile-Specific Treatment Effects Separately by 20% Interval of Honors Enrollment Share



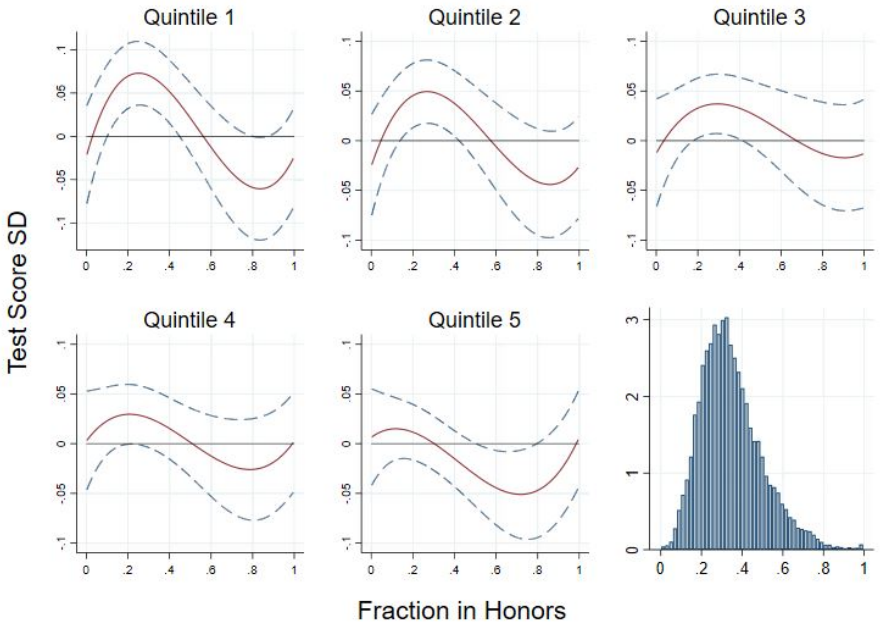
Notes: The first five graphs plot estimates of the function $E[\Delta\overline{Y}_q(f)]$ that maps coursewide honors enrollment fraction into expected standardized test performance by quintile of predicted performance for a specification that interacts indicators for student preparedness quintile with indicators for whether the current course' honors share falls in a particular interval of width 0.2 (with the last two intervals combined due to minimal support). The bottom right graph in each panel displays the density of honors enrollment shares for the chosen sample. 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes. The figures rely on the baseline sample of school-course-year-quintile observations (See Section 3.1 for details).

Figure A.7: Testing Robustness to Alternative Functional Forms for the Treatment Effect Functions $E[\Delta\overline{Y}_q(f)]$: Unrestricted Cubic Specification



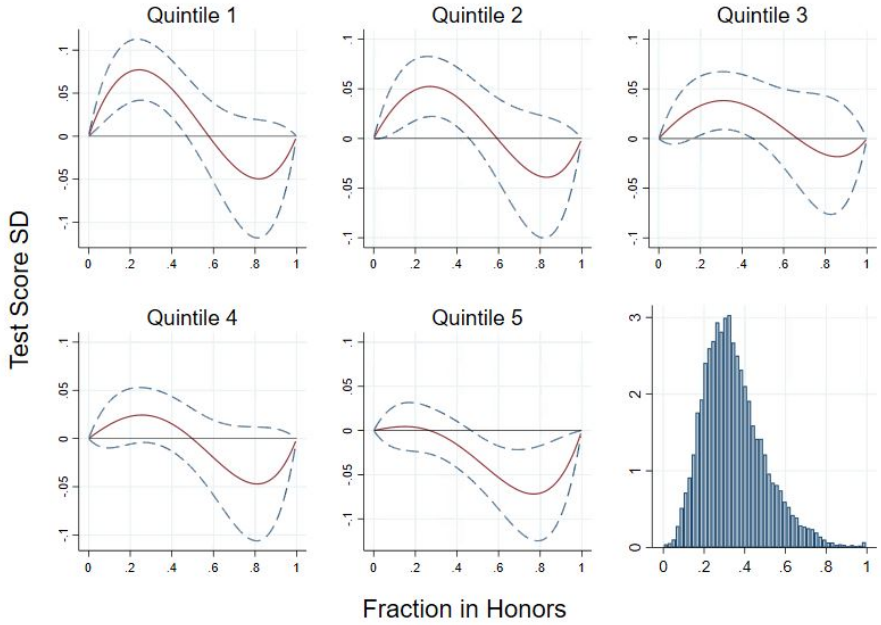
Notes: This figure plots estimates of the function $E[\Delta\overline{Y}_q(f)]$ that maps coursewide honors enrollment fraction into expected standardized test performance by quintile of predicted performance for a specification that does not restrict the value of the treatment effect to be zero at the right end of the unit interval. 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes. The bottom right graph in each panel displays the density of honors enrollment shares for the chosen sample. The figures rely on the baseline sample of school-course-year-quintile observations (See Section 3.1 for details).

Figure A.8: Testing Robustness to Alternative Functional Forms for the Treatment Effect Functions $E[\Delta\overline{Y}_q(f)]$: Discontinuity Permitted at a Zero Honors Enrollment Share



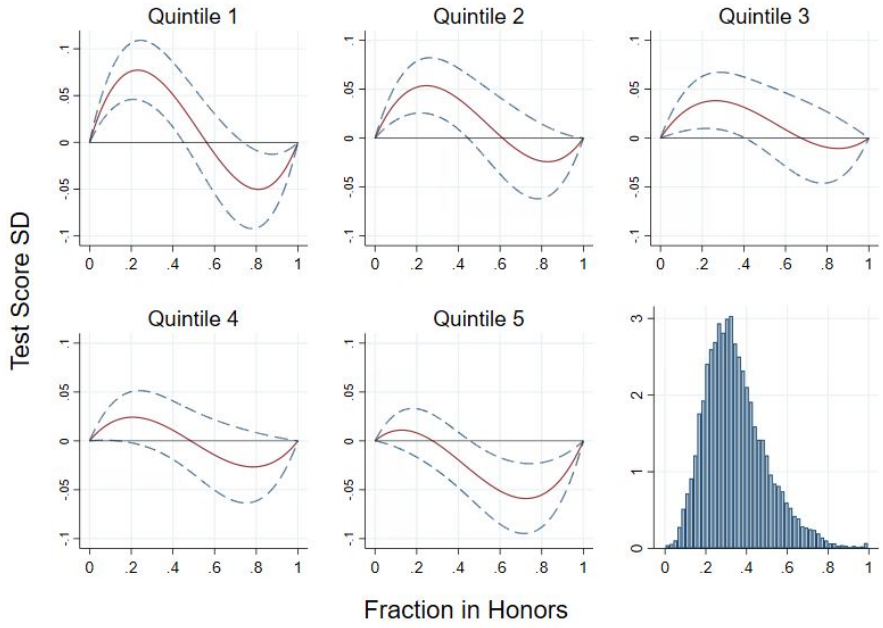
Notes: This figure plots estimates of the function $E[\Delta\overline{Y}_q(f)]$ that maps coursewide honors enrollment fraction into expected standardized test performance by quintile of predicted performance for a specification that includes a separate indicator for whether the course features any tracking. This ensures that predicted values at low enrollment shares are not affected by performance in untracked schools or courses. 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes. The bottom right graph in each panel displays the density of honors enrollment shares for the chosen sample. The figures rely on the baseline sample of school-course-year-quintile observations (See Section 3.1 for details).

Figure A.9: Testing Robustness to Alternative Functional Forms for the Treatment Effect Functions $E[\Delta\overline{Y}_q(f)]$: Restricted Quartic



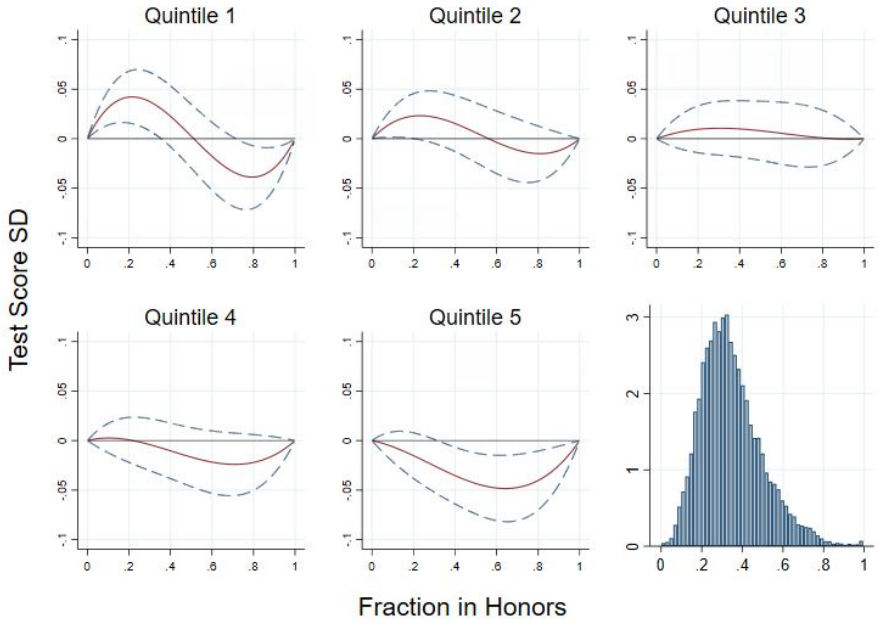
Notes: This figure plots estimates of the function $E[\Delta\overline{Y}_q(f)]$ that maps coursewide honors enrollment fraction into expected standardized test performance by quintile of predicted performance for a restricted quartic specification. We restrict the quartic to take the value 0 at both ends of the unit interval, so that there are three free parameters estimated for each quintile of predicted performance. 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes. The bottom right graph in each panel displays the density of honors enrollment shares for the chosen sample. The figures rely on the baseline sample of school-course-year-quintile observations (See Section 3.1 for details).

Figure A.10: Testing Robustness to Alternative Functional Forms for the Treatment Effect Functions $E[\Delta\overline{Y}_q(f)]$: Using the Share of Honors Classrooms as an Instrument for the Honors Enrollment Share



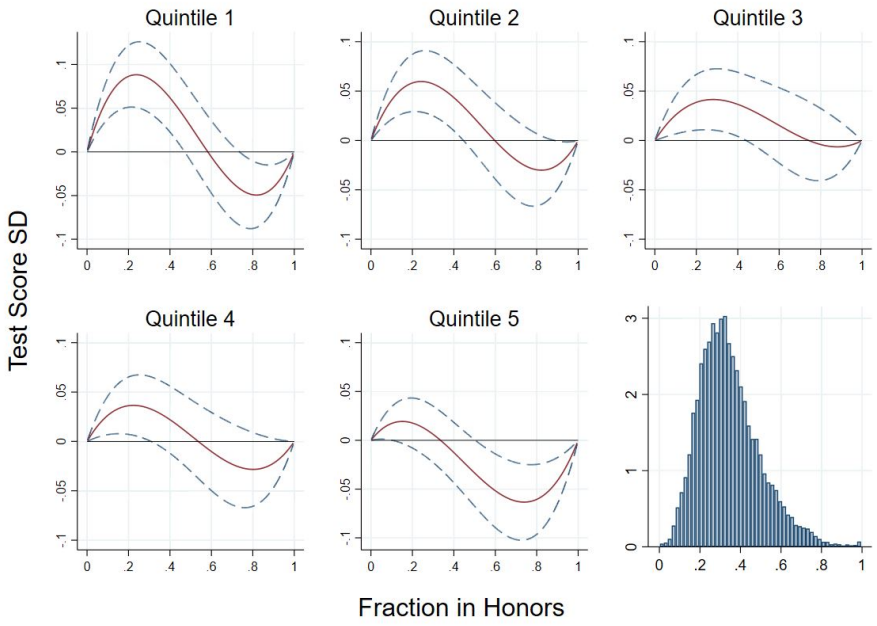
Notes: This figure plots estimates of the function $E[\Delta\overline{Y}_q(f)]$ that maps coursewide honors enrollment fraction into expected standardized test performance by quintile of predicted performance for a specification that instruments the course's share of enrollment in the honors track (and its square) with the course's share of honors classrooms (and its square). 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes. The bottom right graph in each panel displays the density of honors enrollment shares for the chosen sample. The figures rely on the baseline sample of school-course-year-quintile observations (See Section 3.1 for details).

Figure A.11: Testing Robustness to Alternative Functional Forms for the Treatment Effect Functions $E[\Delta\overline{Y}_q(f)]$: Specification Featuring School-Year Fixed Effects



Notes: This figure plots estimates of the function $E[\Delta\overline{Y}_q(f)]$ that maps coursewide honors enrollment fraction into expected standardized test performance by quintile of predicted performance for a specification that augments the baseline specification by including a set of school-year fixed efffects. 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes. The bottom right graph in each panel displays the density of honors enrollment shares for the chosen sample. The figures rely on the baseline sample of school-course-year-quintile observations (See Section 3.1 for details).

Figure A.12: Testing Robustness to Violations of Assumption 1: Specification Featuring a Restricted Sample of School-Courses Featuring More Typical Distributions of Predicted Student Performance Based on Middle School Performance



Notes: This figure plots estimates of the function $E[\Delta\overline{Y}_q(f)]$ that maps coursewide honors enrollment fraction into expected standardized test performance by quintile of predicted performance for the baseline specification but using an alternative sample that restricts the set of school-courses to those where the average student would need to shift their quintile of the preparedness index by less than 1/3 in order for the school-course to match the statewide (uniform) distribution of quintiles. 95% pointwise confidence intervals computed using the delta method are displayed with blue dashes. The bottom right graph in each panel displays the density of honors enrollment shares for the chosen sample.