DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 19-03

Who Benefits from a Smaller Honors Track?

Zach Szlendak
University of Colorado Boulder

October 15, 2019
Revised October 31, 2019
Revised November 1, 2019
Revised November 11, 2019
Revised November 22, 2019

Department of Economics

University of Colorado Boulder
Boulder, Colorado 80309

Who Benefits from a Smaller Honors Track?

November 22, 2019

Zach Szlendak

Current version

Abstract

In 2018, over 13 million American high school students attended high schools with some sort of honors program or ability tracking. However, to date, there is no consensus in the literature on the effects these programs have on academic performance. I show the relative size of honors programs can explain the papers' varied findings. Using data from North Carolina public high schools, I identify the effect of different honors program sizes on test score performance by using variation across schools, within schools across courses, and within schools across time. To address concerns that the gains from honors programs are at the expense of disadvantaged students, I estimate different average treatment effects by quintile of student ability. I find that the optimal honors program size has 20% to 30% of students in it. If all schools switched from their current honors program size to the optimal size, North Carolina high school students would gain an average of 0.02 SDs. For honors programs with more than 35% of students in them, decreasing the honors size leads to a Pareto improvement, by quintile.

1 Research Question and Motivation

Tracking is the process of separating students by ability in order to customize the level of content 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 marginal individuals track choices while several others consider the impact of introducing tracking or removing it entirely.[1] Yet among schools that offer an honors track, there is wide variation both across schools and within schools across courses (documented below) in the share of students that enroll in honors. Motivated by lack of consensus in the optimal honors track size, this paper considers the school's choice of how selective to make its honors track. Specifically, I estimate separate flexible functions mapping a course's fraction in honors into expected standardized test score performance by category of student preparedness. I further show that these functions are sufficient to determine the administrator's optimal choice of honors track size for a typical high school environment where students can self-sort into honors, but where the administrator can adjust the costs of doing so to select their preferred honors track size.

Ex-ante the effects of shrinking the honors program are ambiguous, varying by the type of student, and dependant on the initial size of the honors program. The top students who remain in the honors track experience a faster pace and more capable peers; some marginal students get pushed to a lower track; infra-marginal students in the regular track experience a more rigorous pace and higher ability peers. Expanding the size of honors programs allows more 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. 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 consider reallocating resources between the tracks, obfuscating the effects different types of students experience.

I estimate a baseline specification using an ordinary least squares (OLS) regression in which test scores are regressed on a restricted cubic function of the fraction of students in honors at the school-course in which the student is attending. Separate cubic coefficients are estimated for each quintile of a student preparedness index based on past test scores.[2] I estimate this equation using data from North Carolina, which contains histories of students' past test scores for a large sample of students from 1995 to 2013. The North Carolina

[1] These papers are discussed in greater detail in my literature review discussion at the end of this section

[2] I divide students into quintiles of observed ability by course based on predicted test scores using the students' history of test score performance in mathematics, English, and science.

data features statewide course specific tests in eleven high school courses, of which I choose six.[3] In order to accurately model nonmonotonic effects that vary by student ability type a large data sample is required. This precludes accurate identification through small scale experiments.

To justify interpreting my estimates of varying the honors program size as causal, it is necessary to assume that, conditional on controls, the variation in the honors track size is unrelated to other school and student inputs that may affect test score performance. Notice that by focusing on the school fraction in honors rather than the individual honors choice, I need not assume that each student's track choice is exogenous conditional on the school-course wide cost of enrolling in the honors track. This allows me to sidestep the selection problems associated with individual choice that has been the central focus of the individual effects literature.

None the less, valid identification of the effect of changing the size of honors programs is empirically difficult because honors program size is partially endogenous to school and student characteristics that affect performance, such as unobservably better cohorts driving both the share of students in honors and test score performance. I address this in several ways. First, I limit my sample to schools with typical student distributions and courses to where honors is the only advanced track. I focus on honors tracking instead of Advanced Placement (AP) or International Baccalaureate (IB) tracking because students in both honors and non-honors tracks are taught to the same test in North Carolina, providing a numeraire of educational gains.[4][5] Secondly, rich controls at the school, teacher, family, and student level, including parental educational attainment, school size, teacher experience, education, and test score performance, and student demographics, capture many of the inputs that drive test score performance and the size of the honors track. Thirdly, I employ a school fixed effects specification that examines within school across course and over time variation. Lastly, I use a lagged honors size instrumental variable (IV) specification that eliminates bias from contemporaneous shifts in unobservable cohort quality. The confidence in my results stems from their consistency under alternative specifications that differently weight several sources of variation. If considerable sources of endogeneity bias were to still remain, they would have to have biases with similar magnitude and direction from several different sources of variation in order to produce such results.

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

[4]For AP and IB students are focused on preparation for both the state standardized test and AP or IB test that provides them the opportunity for college credit.

[5]In North Carolina, students in both honors and non-honors have incentive to perform well as state test performance contributes to a students GPA. There is no evidence of grouping at the ceiling of the score range for honors students.

Why might the residual variation of the share of students in honors, after the included controls, be independent of achievement determinants? Administrators or department heads may have idiosyncratic tastes or beliefs on the optimal size of an honors track even conditional on a similar distribution of student attributes, perhaps due to disproportionate pressures from parents or various federal and state educational accountability regime. Variation in beliefs about the optimal size of honors may also be partially driven by the dearth of research on the subject.[6] Secondly, 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, which could affect peer composition and level of instruction in both the honors and non-honors classrooms. Lastly, institutional momentum may resist changes to the share of students in the honors track, even if the current share is sub-optimal. This institutional momentum may take the form of administrator comfort or efforts to limit the number of new classroom preps for teachers.

To show how administrators can select the size of an honors program without assigning students to tracks, I propose a simple mode of students self-sorting. I assume that administrators can alter the share of students in honors by adjusting the costs students face when enrolling in the honors track, creating a default for some students that can be overcome by paying an effort, convenience, or grade cost.[7][8] In practice, while most schools are not explicit about their target size, they are implicitly setting the fraction through policies that affect incentives to enroll or not enroll in honors. These policies include Grade Point Average (GPA) boost of each track,[9] mandatory meetings with counselors before enrolling to either encourage or discourage honors, homework loads in each track, and scheduling convenience of each track. If administrators know the joint distribution of effort costs and observed and unobserved ability, then their choice of enrollment cost determines the expected composition of students in honors.

I find that the highest ability students, quintile 1, most benefit from honors programs that comprise 20-30% of the student body, yielding an increase in test scores of 0.07 SD on average relative to a no tracking alternative. The second quintile exhibits similar but smaller effects as the first, with an average test score gain of about 0.05 standard deviations (SDs)

[6] Conversations with North Carolina administrators and teachers reinforced my belief that their is significant heterogeneity on what is perceived to be the optimal honors size.

[7] Even if all students do not pay the costs, it is sufficient to have students near the margin of tracking into the honors or regular tracks pay the cost.

[8] Including these costs in welfare estimates, similar to Fu and Mehta (2018), would require additional assumptions reducing the validity of my results.

[9] GPA boosts involve adding a numerical value that effectively inflates the letter grades of honors courses when computing the GPA.

for the 20-30% range, but the test score gains for this 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 30-40% of the student body is enrolled in honors. The fourth quintile is relatively unaffected by varying the size of the honors program, but does exhibit small gains of about 0.025 SDs when the share of students in honors is between 20 and 30%. The fifth quintile does not exhibit any statistically significant gains from any exclusiveness and is instead hurt by tracking programs with more than 40% of the student body in them.

When administrators weight the gains of all quintiles equally, honors programs with 20-30% student body enrollment maximize the school's average score, with average gains of 0.04 SDs compared to the absence of an honors track. If all schools switched from their current honors program size to the optimal size, North Carolina high school students would gain an average of 0.02 SDs. The 20-30% range for the share of students in honors still maximizes the administrators problem and delivers sizable gains even with a weighting system that weighs quintiles 1, 2, 3, and 4 at 20%, 40%, 60%, and 80% of quintile 5, respectively. For honors shares greater than 30%, it's likely that the benefit of having more students placed into the honors program gets drowned out by the cost of having both the regular and honors track decrease their average student quality and the level of instruction.

Changing the size of the honors program is a low cost avenue for improvement with potential for sizable lifetime effects. A 0.1 standard deviation (SD) increase in contemporaneous test scores due to teacher performance leads to an increase of annual earnings of at least 1% at the age of 28 (Chetty et al., 2014a,b). If contemporaneous test score gains from the choice of honors track size, then even policies that generate small gains in test score performance may have large lifetime impacts. Furthermore, if the policies affect a large number of students (as is the case in tracking) then these effects can aggregate to very large effects on annual earnings as well as other long run outcomes. Using a back of the envelope calculations from combing the results from Chetty et al. (2014a,b) with my results, if North Carolina high schools changed from their current honors program size to the optimal honors program size for six core courses, then the aggregate increase in earnings at the age of 28 would increase by $44 million annually.

My paper is fundamentally different from the existing literature, as it is the first to look at honors program size in a context where students can self-select their track. My results do have contributions to other strands of the literature and are capable of resolving differential estimates from the literature on the existence of tracking programs. When examining the effect of having tracking programs of unspecified sizes, some papers have found they help the top students and hurt the bottom students (Betts and Shkolnik, 2000; Hoffer, 1992; Argys

et al., 1996; Epple et al., 2002; Fu and Mehta, 2018). Others have found they do not hurt any students (Zimmer, 2003; Figlio and Page, 2002; Duflo et al., 2011; Card and Giuliano, 2016) or have small or insignificant effects (Pischke and Manning, 2006; Lefgren, 2004). My results suggest that these seemingly contradictory results can potentially be reconciled if the different papers feature samples of schools with different mixes of honors program sizes.

Figlio and Page (2002) examined the effect advanced tracks had on high school math performance. To overcome endogeneity in school tracking selection, they employed an IV specification using variation in county policies. While this allowed them to account for students selecting into schools, they did not examine how the size of an advanced track affects outcomes. Additionally, this paper used data from the National Education Longitudinal Study of 1988, which does not state whether students are allowed to self-sort into tracks and often requires economists to infer tracking regimes. I do not have an instrument for school choice and thus do not attempt to assess the impact of school tracking policies on between school sorting. However, I do control for the impacts of between school sorting on student achievement distributions using a rich set of controls capturing past student performance, family characteristics, school characteristics, and a school fixed effects specification. Fu and Mehta (2018) build a structural model that incorporates the administrators choice of the fraction of students to assign to the advanced track. The model permits heterogeneous effects for the tracking schemes that vary with the size of the program in an environment where administrators assign elementary school students to different tracks. The authors are forced to infer the track based on the teachers self report of the quality of the students, which could simply be attributable to sampling error. Duflo et al. (2011) was able to randomly assign ability tracking to elementary schools in Kenya and found increases in test score performance. However, the institutional context differs dramatically from mine, and they do not consider the impacts of changing the size of the advanced track.

A second strand of the literature considers the effect on an individual moving into an honors or gifted track and generally finds that enrolling in advanced tracks improves test scores for the marginal students they consider. My 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. Card and Giuliano (2016) adopts a regression discontinuity design that exploits policies that create advanced elementary school tracks based on thresholds of observable characteristics. Existing attempts at identifying this affect in high school has relied on propensity score matching, which has the potential to amplify omitted variable bias (Hoffer, 1992; Long et al., 2012; Smith and Todd, 2001). My results suggest the impact of honors is not limited to just the marginal students. Students whose past test scores predict they will always enroll in honors or never enroll in honors are still affected.

My work also contributes to the much larger literature considering peer effects on academic achievement. While I do not explicitly isolate peer effects in my model, they are likely to be one of the driving forces for my results. Hanushek et al. (2006); Lefgren (2004) found that having better peers improved outcomes for students across the ability distribution. Mehta et al. (2019) found that improved peer quality increases academic performance through both congnitive and non-cognitive mechanisms, such as study time. Imberman et al. (2012) found similar monotonic peer effects and showed they are not linear. Specifically, they found that the highest ability students were the most sensitive to the quality of their peers. My results are consistent with the possiblity that North Carolina high school students have a similar sensitivity to peer effects as these students. Specifically, my results show 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, as these students are the least sensitive to the peer effects from top students. By employing a structural model and using assumptions about student assignment, Fu and Mehta (2018) was able to separately identify peer effects and found similar results in an elementary school setting. Changing the fraction of student in honors induces different peer effects which differ by the type of student affected. While I do not isolate peer effects in my model, they are likely one of the driving factors for my results.

The remainder of my paper will be structured as follows: Section 2 presents a model of the administrators problem when students self-sort into tracks, Section 3 describes the data, Section 4 lays out my empirical approach, Section 5 reviews the results, Section 6 provides several robustness checks, and Section 7 interprets the findings and concludes.

2 Model

2.1 School's Objective Function

School policy makers have choice over how many honors seats and/or classrooms there are and how students are sorted into honors classrooms. The latter varies from school to school, but most allow students to request a course and track then, if necessary, fill in empty seats using the remainder of the student body. This paper focuses on how the former affects the school production function, specifically how the fraction of students in honors affects educational outcomes when schools allow for some level student or parent choice when sorting into tracks. Schools can influence the fraction directly by setting a hard cap on the number of honors seats or through indirect measures. The former is rare as there appears to be no clear

cap on the size of honors classes in the data.[10] Indirect tools for administrators to set the fraction of students in honors work through altering student incentives to enroll in honors. Schools can affect these incentives by changing the difficulty of each track, the homework loads in each track, the scheduling convenience of each track,[11] the GPA boost of each track, mandatory meetings with counselors who can encourage or discourage a student to enroll in the honors track, and various other policies that affect the costs and rewards of each track. After students sort based on either direct or indirect administrative approaches then the rigor of the classes will be solidified. For this, teachers, administrators, and department heads optimizing for student ability composition will determine the rigor of the honors and non-honors classes based on the size of the honors track. Failure to optimize for student ability accurately will change the enrollment decision of future students. Providing another tool to change the size of the honors track[12].

For the administrator's optimization decision I collapse the administrators direct and indirect tools for setting the fraction of students in honors, f, into a one-dimensional net cost of enrolling in honors class, α_{stj}. Details of this cost are explained with the student's decision in section 2.2. Administrators maximize aggregate student test scores, Y, at school s, in year t, in a course j. Students have different observed ability types, q, and administrators can weight these types differently.

$$\underset{f}{\text{argmax}} \sum_{i=1}^{N} \theta_{q(i)} Y_{istj}(f) \tag{1}$$

The weights, θ_q, sum to one and capture administrators' preferences for which students improve. The weights allow for administrators to prioritize academic growth for different observed types in order to satisfy local, state, and federal educational objectives, such as no child left behind, satiate different parents, or match their preferences for different types of students.

2.2 Test Score Production Function

Educational production, Y_{istj}, depends on whether the student is enrolled in the honors class, $h(q_{istj}, \epsilon_{istj}|\vec{q}_h, \vec{\epsilon}_h)$, or the regular class, $r(q_{istj}, \epsilon_{istj}|\vec{q}_r, \vec{\epsilon}_r)$. The students choice of track is represented by the indicator $\mathbb{1}(i \in h)$, with 1 signifying enrollment in honors. Some students

[10] North Carolina has caps on the classroom size, but they do not appear to influence classroom size significantly.

[11] Schedules are generally made by administrators then students either choose or request classes from available options.

[12] Enrollment in the honors track is on a per course basis and students request their track in the proceeding Spring or Summer. As a result, the response to the level of difficulty in a class happens the next year.

may benefit more from an small honors than one that accepts a large share of the student body. To account for this, the efficacy of these classes will depend on the observed ability composition, $\vec{q}_r$ & $\vec{q}_h$, and the unobserved ability composition, $\vec{\epsilon}_r$ & $\vec{\epsilon}_h$, of those classes. Students' production in the classroom will be affected by how the material matches with their ability and how other peer effects interact with their own type. Hence, the efficacy of different tracks will have heterogeneous effects on students depending on how the student's observed ability type q_{istj}, and the student's unobserved ability type, ϵ_{istj}, are complimented by the level of instruction and the peer effects in the chosen track. Other observed school, peer, family, and individual inputs, X^O_{istj}, and unobserved school, peer, family, and individual inputs, X^U_{istj}, will contribute to the students education production. Observed and unobserved inputs to the education production function not related to tracking, $X^O_{istj}\beta^O + X^U_{istj}\beta^U$, will remain unspecified at this point as they are irrelevant to the school's tracking decision. The student's educational production function is:

$$Y_{istj} = [1 - \mathbb{1}(i \in h)]r(q_{istj}, \epsilon_{istj}|\vec{q}_r, \vec{\epsilon}_r) + \mathbb{1}(i \in h)h(q_{istj}, \epsilon_{istj}|\vec{q}_h, \vec{\epsilon}_h) + X^O_{istj}\beta^O + X^U_{istj}\beta^U + \mu_{istj}$$

Honors and regular classes are tested with the same standardized test. Administrator, parent, and student preferences for high scores will help ensure the curriculum for both tracks will be similar. The main differences in the two tracks will be the level and depth of the instruction and the different peer effects. It is possible that teachers of different quality are allocated differently to each track. Teacher assignment systems are mediated through the treatment effect. Generally senior teachers get assigned to honors tracks, but there may be an assignment decision based on unobserved ability or specialized human capital (Cook and Mansfield, 2016). This is an additional dimension of choice that hasn't been modeled separately and is instead incorporated through the estimates. The heterogeneous effect of honors and regular classes will mean that students will choose which class to enroll in based on their gains from each track and their tastes. I will isolate these effects from other student, school, and family effects with Assumption 1. Specifically,

Assumption 1. *$r(\cdot)$, $h(\cdot)$, $X^O_{istj}\beta^O$, $X^U_{istj}\beta^U$, and μ_{istj} are additively separable.*

This is a less restrictive assumption than often used in the literature as it makes no assumptions about observed quality, q, or unobserved quality, ϵ, having constant effects on student outcomes regardless of track. Propensity score matching papers that rely on observable similairites of students in each track assume the honors and regular production functions have the same affect for ϵ, specifically $h(q, \epsilon) = h_1(q) + b(\epsilon)$ and $r(q, \epsilon) = r_1(q) + b(\epsilon)$. If this were the case, then it would be possible to look at the impact of different honors policies without worrying about the effect of unobservable characteristics. Under this analysis, there

would be limited gains from students unobserved ability complementing the level of the class and the highest unobserved ability students would benefit from honors just as much as the lowest unobserved ability students, conditional on observed quality. In terms of applicability of the model, most high school honors programs do not sort purely on observed quality, q.

Past performance is an imperfect estimate of ability. This may be worsened in the ninth and tenth grades, the grades most students enroll in the courses in my model, due to potential changes in a student's motivation and work ethic. Some observably high quality students may not plan on attending college and may focus their efforts more on employment while in high school. Some observably low quality students may increase their effort in hopes of attending colleges, as colleges use high school transcripts for admittance decisions. Factors like drive and work ethic affect student academic gains from each track. Generally more driven, higher ability students benefit more from the honors track. As a result, variation in those factors that affect student's gains will be used when sorting based on academic gains from complementarity.

Students also have factors that affect sorting decisions outside of academic gains from complementarity. Specifically, students face different costs from each track. As mentioned in section 2.1, administrators have the ability to change the cost students face when enrolling in different tracks. Some options to affect this cost, such as changing homework loads, directly affect academic gains and are partially captured by $h(\cdot)$ and $r(\cdot)$, reducing the sorting effects these costs have. Other options, such as changing GPA rewards for the honors track, may indirectly affect learning. GPA rewards could incentivize students to change how hard they work, but are likely not changing $h(\cdot)$ and $r(\cdot)$ significantly. Because all students have to take a track, I only examine the difference in the costs of the different tracks. I will introduce a cost, c_{istj}, which captures the difference between the student's idiosyncratic cost of the honors and regular classes and a cost, α_{stj}, which expresses the shared cohort difference between the cost of the honors and regular classes. Positive values of c_{istj} and α_{stj} indicate that the honors class has a larger net cost to the individual and cohort respectively, while a negative value indicates that the honors class has a smaller net cost. While time and effort are usually greater for honors classes, GPA boosts in honors class give the administration a tool to lower the cost differential for the two tracks.

The shared cohort cost difference, $\alpha^*_{stj}(f)$, is the cost that will lead to a fraction of f students enrolling in the honors track. This is the choice variable for administrators and is exogenous for the student's choice of track, conditional on f. The sorting decision for each student, combined with Assumption 2, will cause each fraction f to yield the same sorting outcome.

Assumption 2. *The joint distribution of q, ϵ, and c, $g(q, \epsilon, c)$, is the same for all cohorts.*

Given a shared cohort cost that yields a fraction f of students in honors, $\alpha^*_{stj}(f)$, the student's sorting decision is given as:

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

where $\alpha^*_{stj}(f)$ is s.t.

$$\iiint_{A(f)} g(q, \epsilon, c) dq d\epsilon dc = f.$$

$A(f)$ is the restricted set of $\{q, \epsilon, c\}$ s.t. $h(q_{istj}, \epsilon_{istj}|\vec{q}_h, \vec{\epsilon}_h) - r(q_{istj}, \epsilon_{istj}|\vec{q}_r, \vec{\epsilon}_r) - c_{istj} - \alpha^*_{stj}(f) > 0$.

I abstract away form finding the sorting equilibrium and instead focus on average effects for each observed type. Assumption 2 combined with the sorting decision makes f a sufficient statistic for each sorting equilibria. This is because Assumption 2 and the sorting decision are sufficient to create an ordinal ranking of students that, when split, creates sets for honors and regular track students that have a 1:1 mapping to f. The ordinal ranking reduces the set of equilibria outcomes to the number of values of f. Each producing the same sorting outcome every time. When administrators do decide what fraction of the student body should be in honors, they assume that students will sort based on observed and unobserved characteristics. Students of a given observed quality type will have a probability of sorting into honors equal to $P(d^h = 1|q_{istj} = q, f)$. Where

$$P(d^h = 1|q_{istj} = q, f) = \iint_{A_q(f)} g(q_{istj}, \epsilon, c) d\epsilon dc \tag{2}$$

and $A_q(f)$ is a subset of $A(f)$ such that $q_{istj} = q$. Once sorted, a student of a given observed quality type will have expected gains given by

$$E_{\epsilon \in h}[h(q_{istj}, \epsilon|f)] = \iint_{A_q(f)} h(q_{istj}, \epsilon|f) g(q_{istj}, \epsilon, c) d\epsilon dc \text{ and} \tag{3}$$

$$E_{\epsilon \in r}[r(q_{istj}, \epsilon|f)] = \iint_{A^c_q(f)} r(q_{istj}, \epsilon|f) g(q_{istj}, \epsilon, c) d\epsilon dc \tag{4}$$

The inability to observe student unobservable quality means administrators have to act on the expected effect of honors and regular classes. If I assumed that unobserved ability has a constant effect on outcomes independent of track, then administrators could recover

the individual $h(\cdot)$ and $r(\cdot)$ functions. However, this is likely not the case as an unobservably high ability student with poor past performance would likely gain significantly more from enrollment in the honors track than an unobservably low ability student with the same past performance. When making policy decisions with regards to the size of an honors program, administrators can ignore observed and unobserved inputs not part of the tracking functions because they are separable from the administrators choice variable as per assumption 1. The administrator's decision will not change if their knowledge is limited to a collective expected treatment effect,

$$E_{q(i)}[Y_{istj}|f] = P(d^h = 1|q_{istj}, f)E_{\epsilon\in h}[h(q_{istj}, \epsilon|f)] + [1 - P(d^h = 1|q_{istj}, f)]E_{\epsilon\in r}[r(q_{istj}, \epsilon|f)] + +X^O_{istj}\beta^O + X^U_{istj}\beta^U \quad (5)$$

All schools are restricted to have the same treatment effect for each value of f. This is not ideal, but perfect identification of $E[Y_{istj}]$ is intractable, as there is insufficient data to have a separate treatment effect for each $\{\vec{q}, \vec{\epsilon}\}$. A small percentage of schools can be dropped to limit the sample to schools that have a similar composition of students, more details about this are in Section 4. The administrators problem, equation (1), can now be simplified to

$$\underset{f}{\text{argmax}} \sum_{i=1}^{N} \theta_{q(i)} E_{q(i)}[Y_{istj}|f] = \sum_{q=1}^{Q}\sum_{i=1}^{N_q} \theta_{q(i)} E_{q(i)}[Y_{istj}|f] \quad (6)$$

Where N_q is the number of students of each observed quality type. Let W_q be the ratio of students of each observed quality type at a given school. Equation (2) can be rewritten as

$$\underset{f}{\text{argmax}} \sum_{q=1}^{Q} \theta_q W_q E_{q(i)}[Y_{istj}|f] \quad (7)$$

The administrators problem to maximize scores through the exclusivity of honors is a one dimensional problem with f being the choice variable. Administrators do not need to worry about students self-sorting. The choice of f is made conditional on students sorting based on observed ability, unobserved ability, and preferences. The effects that the administrators problem captures are not just students who have their sorting decision changed through changes to f. The honors and regular functions capture how changing f changes the peer effects and level of instruction.

3 Data & Background

My data covers all North Carolina public schools from 1995-2013. The data is panel, but some variables are not sampled every year. Student data is from the third grade through

high school and covers all students in North Carolina Public schools. Student-level data includes economic disadvantage status, census block group, race, sex, parental income, gifted status, disability status, parental education, and academic performance measured by grades and test scores. Students in elementary and middle school are tested by state end-of-grade math, science, and reading tests. High school student data is limited to 11 subjects: Algebra 1, Algebra 2, Biology, Chemistry, Civics, Econ Law & Politics, English 1, Geometry, Physics, Physical Science, and US History. Students are tested by a state end-of-course exam in each of the 11 previously mentioned high school classes. Student performance on these exams contributes to at least 20% of the students course grade, providing students incentive to perform well. Appendix Figure A.1 looks at the statewide distribution of student scores of the courses in my final specification for the year 2006.[13] There is little reason to be concerned about student scores being grouped at the upper or lower limit of the score range. Classroom data includes period, size, and track. Teacher data includes teacher experience, educational level, educational institution, course load, and certification(s). School data includes student counts by race and sex, student counts by economic status, school achievement level, school safety metrics, and the percent of students above various thresholds. Student, family, classroom, teacher, and school data can be linked.

The share of students in honors classes can be measured using either student level or classroom level data. I will use the classroom level data for two reasons. First, using student reported honors shares could bias my results upwards. If a high quality student's observation was dropped, then that would likely decrease the share of students in honors observed and average performance in that subject. If a low quality student's observation was dropped, then that would likely increase the observed share of students in honors and average test score performance. Second, students can misreport their track. This is seen by a small fraction of students reporting tracks that are inconsistent with what the classroom data and the students peers, including tracks not offered in that course. Table 1 examines what tracking options schools offer in each course for school-year-courses (cohorts) with at least 30 student observations. For most cohorts there exists an honors program, but remedial programs are rare. The remedial track generally has a very small portion of the student body in it when it exists. Figure A.2 examines the share of students in honors of the remedial track for the subset of school-year-courses where the remedial track exists.

Due to the presence of AP offerings in US History and Physics, they will be excluded from my analysis. By focusing on AP test preparation, the treatment for tracking in these subjects includes different curricula. Due to concerns over different curricula, IB schools are

[13]More years are available by request from the author. No course-year in my sample exhibits bunching around a ceiling or floor.

excluded in all subjects. Civics and Economics, Law, & Politics were sampled infrequently or had a large variety of courses that take those tests and will be excluded. Algebra 1 has low honors frequency for high schools largely because many students take it in middle school. As a result, the quality distribution of students that take it in high school is abnormal and using a one dimensional metric for the effects of honors sizes would be insufficient. Hence, Algebra 1 will be excluded from my analysis. Appendix Figure A.3 examines the share of students in honors for the courses in my final sample for school-year-courses with honors programs. Most of the support is in the range of 0.1 to 0.6 for most courses. Chemistry has a higher fraction of students in honors on average than the other classes I will be examining.

For all of my analysis, I will assign students observed quality types by predicting statewide quintiles, with quintile 1 being the students predicted to perform the best. The quintile prediction is based on the past history of student test scores.[14] The quintiles for each student are course specific and based on statewide performance. A student who has excelled in past math and science tests, but struggled on past English tests, may be in a high predicted quintile for Science, Technology, Engineering, and Math (STEM) subjects and a low predicted quintile for the humanities. I will refer to these statewide predicted predicted quintiles as just quintiles for the rest of the paper. For some figures I use within school student rankings for quintiles. For these cases I clarify that the quintiles are within-school.

I restrict my sample to schools with similar distributions of student characteristics. This restriction enables me to collapse the exclusivity of honors and the corresponding peer effects and level of instruction to a one dimension metric. In my model, this restriction corresponds with Assumption 2. Appendix Figure A.4 looks at how many quintiles students would need to shift on average in order for a school to achieve a uniform distribution of predicted quintiles. By limiting my sample to schools where the average number of shifts is less than or equal to one half, I will drop about 25% of my observations. This will limit the external validity to very good and very bad schools. Appendix Figure A.5 is the histograms of the six schools with the aforementioned metric closest to and less than one half. Table 2 has the school-course-year averages for high schools after dropping atypical schools, by honors program size. Besides smaller schools being more likely to not have honors programs, there is little reason to believe that schools with larger or smaller honors programs are significantly different. However, the majority of schools in North Carolina are fairly uniform when it comes the distribution of their student quality. For application of my results to contexts outside of North Carolina, the state ranks in the middle of US states for educational performance.[15]

[14] Specifically, for each course $PredictedScore_{istj} = english7_{istj}\beta_j^1 + math7_{istj}\beta_j^2 + english8_{istj}\beta_j^3 + math8_{istj}\beta_j^4 + \epsilon_{istj}$. Results are robust to the inclusion of science, however science test scores have fewer observations.

[15] Education Rankings (2019)-*US News* https://www.usnews.com/news/best-states/rankings/

As a robustness check, I will also include a specification where the above metric for the spread of student quality is less than one third. Appendix Figure A.6 in Appendix A is a histogram of the six school-courses with the average number of shifts required for a uniform distribution closest to a third.

Figure 1 plots the average test score performance, in statewide SDs, for bins of the share of students in honors by quintile. By quintile, it appears that most types of students' average test scores are largest when the share of students in honors is around 40%. This could be due to a causal effect or it could be due to correlation with other school and student factors. The high variability of average scores for shares of honors above 65% is due to the lack of support for this range of the data.

Students in top quintiles enroll in honors at a higher rate than students in other quintiles. Figure 2 plots the average honors enrollment rate for bins of the share of students in honors by quintile. Perfect sorting on observables would result with a linear trend starting with a quintile specific honors enrollment rate of 0 at a school-year-course with a share of students in honors of 0; then end with a quintile specific honors enrollment rate of 1 at a school-year-course share of students in honors of 0.2 for quintile 1. For the second quintile, the quintile specific honors enrollment rate would be 0 until the school-year-course share of students in honors reached 0.2, at this point there would be a linear trend until the quintile specific honors enrollment rate is 1 for a school-year-course share of students in honors of 0.4; etc. The difference between these perfect sorting plots and the actual sorting plots shows that there is still a lot of assignment on factors other than the past history of student performance.[16] Some students in the bottom quintile of their school sign up for honors when there are few seats and some students in the top quintile of their school refuse to sign up for honors even when half the seats in a school-year-course are in an honors class. This is due to students having high or low unobserved ability, students having a high or low cost for honors enrollment, and potentially administrators filling seats. For quintiles 2 and 3 these unobserved sorting factors play a large role in enrollment as they both have significant enrollment rates for all shares of students in honors. The sixth cell in Figure 2 shows the support of the data along the share of students in honors.

Most administrative tools to change the share of students in the honors track cannot be observed in the data. Some, such as increased rigor or homework loads, cannot be separately identified. For example, when including school, year, and course FEs, an increase of one hour per week spent on homework in an honors program is associated with a 0.5 percentage

education

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

point decrease in the share of students enrolled in honors. This could be due to students not enrolling in response to an increased cost, the marginal student spending less time on homework, or other factors. There is one administrative tool that is observed and separately identifiable, GPA boosts.[17] Increasing the honors GPA boost by one point increases the share of students in honors by 12 percentage points.[18] Unfortunately there are insufficient observations with GPA boosts to use it as an instrument. However, it does indicate that the administrative tools to change the costs have a significant impact on the share of students in honors.

4 Empirical Approach

4.1 Primary and Two Alternate Specifications

All my specifications are quintile specific conditional expectations functions, $E_{q(i)}[Y_{stji}|f]$, of test score outcomes conditional on the share of students in the honors track, mirroring equation 5 in Section 2.2. Quintiles of student preparedness, used as proxies for observed student ability, are measured by a regression index from a regression of current test scores on past test score.[19] Since all controls at the individual level are linear and the share of students in honors does not vary within a school-year-course-quintile, the data is collapsed down to the school-year-course-quintile level. The average score, $\bar{Y}_{stjq}$, for school s, in year t, in course j, and in observed quintile q, is assumed to depend on a function of the share of students in honors, $h^q(z_{stj})$, other observable characteristics X_{stjq}, a vector of fixed effects (FEs) at various levels Γ_{stjq}, and an error term clustered at the school level μ_{stjq}. All my estimating equations are particular cases of the general form

$$\bar{Y}_{stjq} = h^q(z_{stj}) + X_{stjq}\beta^X + \Gamma_{stjq}\beta^\Gamma + \mu_{stjq} \tag{8}$$

Suppose my functional form for $h^q(z_{stj})$ is the true effect for the honors and regular tracks and that $X_{stjq}\beta^X + \Gamma_{stjq}\beta^\Gamma$ captures the observed component of equation 5 in Section 2.2. By Assumption 1, the unobserved component of equation 5 is independent of $h^q(z_{stj})$ and my observable controls. Hence it is captured by μ_{stjq}. Without further assumptions, my estimating equation can be derived from the production equation, equation 5.

The effect of varying the share of students in honors is captured by $h^q(z_{stj})$. It is a cubic in which coefficients are interacted with quintile to enable heterogeneous effects by levels of

[17] The honors GPA boost works by adding points to a student's numerical value associated with a letter grade. A one point boost would make a "B" in an honors class contribute 4 points towards the students GPA instead of the standard 3.

[18] The regression ran was $shareinhonors_{stj} = GPA\_boost_{stj} + \alpha_j + \delta_t + \zeta_s + \mu_{stj}$. The coefficient on $GPAboost$ had a value of 0.118 and a SE of 0.0218

[19] More detail on these can be found in Section 3.

student preparedness:

$$h^q(z_{stj}) = \gamma_q^{lin} z_{stj} + \gamma_q^{sq} z_{stj}^2 - (\gamma_q^{lin} + \gamma_q^{sq}) z_{stj}^3 \tag{9}$$

The coefficients on equation 9 restrict the treatment effect to be the same when placing zero students in honors classes and when placing all the students in honors classes, since both scenarios arguably represent an absence of tracking.[20] This restriction to the model improves precision. I also present results from an unrestricted cubic specification as a robustness check in Section 6.

Exogenous across-school variation in the share of students in honors stems from a list of sources. However, there may exist across school variation in student, school, and teacher attributes that are correlated with the share of honors that will drive the estimates of the $\vec{\gamma}^{lin} = \{\gamma_1^{lin}, ..., \gamma_5^{lin}\}$ and $\vec{\gamma}^{sq} = \{\gamma_1^{sq}, ..., \gamma_5^{sq}\}$ to deviate from their true values. Controls are included to mitigate concerns about potential bias. The vector X_{stjq} contains student ability status, demographics, and past performance, family socioeconomic and education status indicators, proxies for teacher quality, and school demographic, size, and accountability measures. These controls are discussed in Section 3 and are designed to absorb influences that drive or respond to the share of students in honors. For Γ_{stjq}, year-course-quintile fixed effects are included in my primary specification. Thus for my baseline results to be valid, one must assume that $E[\mu_{stjq} z_{stj} | X_{stjq}\beta^X, \Gamma_{stjq}\beta^\Gamma] = 0$, $E[\mu_{stjq} z_{stj}^2 | X_{stjq}\beta^X, \Gamma_{stjq}\beta^\Gamma] = 0$, and $E[\mu_{stjq} z_{stj}^3 | X_{stjq}\beta^X, \Gamma_{stjq}\beta^\Gamma]$=0. The identification of $\vec{\gamma}^{lin}$ and $\vec{\gamma}^{sq}$, the parameters of interest, relies on three different sources of variation in the share of students in honors, z_{stj}: 1) across-school variation, 2) within-school-across-year variation, and 3) within-school-year-across-course variation.

Without controls across-school variation is prone to bias due to students' school selection. Robust student, family, teacher, and school level controls help to mitigate these biases. When examining teacher value added, Chetty, Friedman, and Rockoff argued that similar controls were sufficient to ease concerns about unobservably better students driving certain teachers to have better growth due to sorting (Chetty et al., 2014a). These controls are likely insufficient to identify tracking effects at the individual level, but it is not necessary to due so. Note that variation in unobserved student characteristics within a preparedness quintile that predict their willingness to take the honors track do not generate bias. This is because every student has to take some track, so positive biases from unobservably superior students

[20] While AP and IB classes teach to a different curriculum, honors and regular classes teach to the same state standardized test that contributes to student grades. As a result three of the largest effects from students tracking into honors, peer effects, allocating teachers between tracks, and specialized instruction, are the same when the fraction of students in honors is equal to zero or one. There may be other small effects due to confidence from the track name or curricular differences.

that take the honors track are offset by negative biases from unobservably inferior students. Having the biases cancel out is why it is easier to identify the quintile effect of tracking than the individual effect. The existence of students sorting into the honors track based on their unobserved quality is insufficient to bias my results. Specifically, $E[\epsilon_{istj}Y_{istj}|X_{stjq}, \Gamma_{stjq}] \neq 0 \nRightarrow E[\mu_{stjq}z_{stj}|X_{stjq}, \Gamma_{stjq}] \neq 0$. For utilizing across-school variation, the set of fixed effects, Γ_{stjq}, cannot include school fixed effects.

One might still be concerned that administrators of high quality prefer a certain size of the honors track, average unobserved teacher quality influences the share of students in honors, or that a school's average unobserved student quality is driving total enrollment in honors. That is why I have a second specification in which I augment Γ_{stjq} by including school fixed effects. While this purges multiple sources of bias in the share of students in honors, it also removes multiple sources of exogenous across-school variation in the share of students in honors. Across school exgoenous variation purged includes idiosyncratic tastes in administrator preferences and beliefs, institutional momentum, and differences in cohort size that change the natural set of honors offerings. Nonetheless, there are several sources of exogenous variation in the share of students in honors within a school due administrator preferences and tastes that vary by course or over time, institutional momentum that varies by course, teachers trying to minimize the number of course-track class preps they have, policies that change across time, and small differences in cohort size that influence the natural set of honors. Unobserved cohort quality would only bias the results if the unobserved quality varied across course or across year.

My primary specification will be an ordinary least squares (OLS) specification without school fixed effects in order to capture all three forms of exogenous variation. It will be aided by having strong controls. All teacher, parental, and family controls will be interacted with course to allow for heterogeneous impacts by course. Past student performance will be interacted with course and quintile to allow for heterogeneous impact by course and to not impose a single linear effect for past test score performance for the entirety of the student preparedness distribution. Teacher controls are at the school level so that the assignment of teacher quality to different tracks is captured as part of the treatment. Given the high level of freedom administrators have with respect to assigning teachers to tracks, I assume they optimize using local knowledge on teacher quality across tracks. Controlling for teacher quality at the classroom level would remove this optimization that maybe considered when selecting the share of students in honors.

For proper identification of $h^q(z_{stj})$, it is necessary average student level characteristics that affect both the share of students in honors and test scores are captured by predicted test scores and observable student, family, teacher, and school characteristics. If these controls

are insufficient, then unobserved quality will likely influence both the share of students in honors and the outcomes for the cohort. Including a school fixed effects specification will be effective at preventing bias if the unobserved quality of students is constant across courses and time. However, these fixed effects also remove many sources of exogenous variation. To address unobserved cohort quality varying across year and driving the share of students in honors, I include a third specification that instruments the current share of students in honors with the previous share.

To remove bias due to a cohort being unobservably better or worse I instrument the current share of students in honors, z_{stj}, with the previous share of students in honors, $z_{s,t-1,j}$, at the school-course. The controls and the set of FEs are the same for this specification as my primary OLS specification. The second stage of the IV regression has the same $h^q(z_{stj})$ as the OLS specification. However the observed share of students in honors z_{stj}, and its polynomials, are replaced with the predicted values from the first stage, $\hat{z}_{stj}$, and its polynomials. Specifically, for the second stage

$$h^q(\hat{z}_{stj}) = \gamma_q^{IVlin}\hat{z}_{stj} + \gamma_q^{IVsq}\hat{z}_{stj}^2 - (\gamma_q^{IVlin} + \gamma_q^{IVsq})\hat{z}_{stj}^3 \tag{10}$$

The first stage regression captures variation in z_{stj} due to institutional momentum, administrator preferences, and department head tastes. These sources of exogenous variation are relatively invariant across time satisfying the relevance restriction. Unobserved quality of a single cohort is eliminated in this IV specification even if the unobserved quality varies by course. The exclusion restriction for this IV specification requires that the past share of students in honors affects current test scores only through the current share of honors conditional on controls. This could be invalid if there were an omitted variable, such as unobserved teacher quality, that drives both the past and present share of students in honors and test score outcomes.

Observations are weighted by the share of the students at the school-year-course that are in each quintile, weighting all school-year-courses equally. A weighting scheme based on the number of students rather than shares would prioritize the efficacy of administrators' actions at large schools over other schools. Given I am interested in identifying the school-course average partial effect and not the population partial effect, this weighting scheme is preferred. In addition to clustering my standard errors, my weighting scheme acts as a correction for school-size-related heteroskedasticity in the error. As per the recommendations of (Solon et al., 2015), I have also rerun my primary specifications with equal weighting placed on each student. Point estimates and standard errors are similar for the different weighting schemes[21].

[21]These specifications are available upon request from the author.

4.2 Other Specifications

In Section 6 I will employ several other specifications to confirm my results are not driven by specification or bias. To confirm that my results are not driven by assumptions created through the functional form of my specification, I employ two specifications that relax the assumptions placed on $h^q(z_{stj})$. Firstly, I employ a specification that does not restrict the treatment effect to be the same at zero and one, such that

$$h^q(z_{stj}) = \gamma_q^{lin} z_{stj} + \gamma_q^{sq} z_{stj}^2 + \gamma_q^{cb} z_{stj}^3 \tag{11}$$

This specification will have decreased precision compared to my primary specification, especially when the share of students is greater than 65% as there are limited observations. My specification also creates the assumption that there is not a discontinuity when the share of students in honors is zero. To address this I will include a specification that allows for this discontinuity. Under this specification, $h^q(z_{stj})$ becomes

$$h^q(z_{stjq}) = \gamma_q^{lin} z_{stj} + \gamma_q^{sq} z_{stj}^2 - (\gamma_q^{lin} + \gamma_q^{sq}) z_{stj}^3 + \gamma_q^{indicator} \mathbb{1}_{(z_{stj} \in (0,1))} \tag{12}$$

Where $\mathbb{1}_{(z_{stj} \in (0,1))}$ is equal to 1 if there is honors tracking and 0 if there is not. This specification will also suffer from decreased precision as it now needs to identify more coefficients per quintile. However, it will provide a good robustness check to see if school-courses where there exist an honors program are substantially different from those that lack honors programs.

I employ two additional specifications aimed at showing my results are not driven by unobserved cohort quality. First is another instrumental variable approach that uses the share of classrooms that are honors classrooms as an instrument for the share of students in honors. If administrators set their course and track offerings based on past student test score performance, then this specification would not be biased by unobserved cohort quality as it would not influence the share of classes that are honors. The first stage of this specification comes from more honors classroom options decreasing the scheduling costs for students trying to enroll in honors and potentially signalling that administrators are encouraging students to enroll in honors programs. Hence, when students meet with academic counselors, they would be encouraged more than normal to enroll in honors. I instrument for the class share as opposed to using it as an explanatory variable because a class share metric is less capable of examining how honors programs that are smaller affect outcomes. The class share is just one contributor to the fraction of students in the honors track. For example, the effects of optimizing level of instruction and peer quality can vary for the same class share if the class size is not constant across tracks. Although this specification is less powerful than my primary specification, its potential to purge endogeneity from unobserved student quality provides a good robustness check.

The other specification to address unobserved cohort quality varying across year and driving the share of students in honors, I also include a third specification that has school-year fixed effects in Γ_{stjq}. These fixed effects eliminate any variation in unobserved cohort quality that is constant across all courses. Remaining variation in the share of students in honors comes from across-course-within-school sources. The potential endogenous variation remaining after the inclusion of school-year fixed effects comes from across course unobserved student quality and unobserved teacher characteristics that may drive the share of students in honors. While the endogenous variation from these sources is likely small or negligible, the inclusion of school-year fixed effects has greatly reduced the amount of exogenous variation left to identify $h^q(z_{stj})$. As a result, it is possible that my results may be biased as the endogenous variation left may be a significant proportion of the remaining variation in the share of students in honors.

Unfortunately, I cannot prove that any one specification is free from endogenous variation in the share of students in honors. However, each specification eliminates certain sources of endogenous variation in the share of students in honors. Unless all the remaining sources of endogeneity produce the same bias, relative to the remaining exogenous variation in each specification, then it is unlikely that my results are due to bias.

5 Results

5.1 Quintile Treatment Effects

Figure 3 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. The red lines in Figures 3 through 5 display predicted values from several specifications of the impact of different honors program sizes on student outcomes separated by student preparedness quintile. Dashed blue lines indicate the upper and lower bounds on 95% confidence intervals (CI) that were created using the Delta Method.[22] All figures present results relative to no tracking, which has been normalized to zero. The bottom right cell in each figure displays the support of the honors share distribution for school-year-courses that have honors. There is limited support for honors programs with shares greater than 65% and between 0 and 15%.[23] Therefore the results from these ranges are driven by functional form and should be viewed skeptically. Table 3 provides linear, quadratic, and cubic coefficients from different specifications in this section as well as in Section 6. The imprecision of the estimates in Figure 3 necessitates the more restrictive cubic form from

[22] All my specifications will use standard errors clustered at the school level.

[23] The gradient in the delta method is equal to zero when the share of students is one or zero.

equation 9, which Identifies two parameters per quintile instead of 4. Nonetheless, one can observe a general patter of top student gaining under smaller honors programs and bottom students losing under larger honors programs.

Figure 4 presents my preferred cubic specification associated with equation 9, that yields more precise estimates. Top students benefit significantly from honors programs with fewer than 30% of students in them, gaining about 7% an SD in state test score performance (about the same amount as switching from the median teacher to a 76^{th} percentile teacher (Chetty et al., 2014a)). However, these gains quickly disappear as the honors program increases in size. This is likely due to the dilution in student quality for the honors program, as students in quintile 1 are likely to enroll in honors if there is at least 30% of the cohort in honors. These results are consistent with the peer effect literature. Imberman et al. (2012) found that high achieving students were especially sensitive to peer effects, potentially justifying why quintile 1 experiences such a sharp decrease as the share of students in honors is increased. Specifically, for honors programs with at least 30% of the cohort in it, those from quintile 1 who are in honors (most of them) have their peer quality decrease as the share of students in honors increase.

Quintiles 2 and 3 benefit the most from smaller honors programs with less than 40% of the cohort in them with average quintile gains of about 0.05 SDs and 0.04 SDs. Even though seemingly similar students in these quintiles frequently self-sort into different tracks. In particular the share of honors in quintiles 2 and 3 for honors programs with around 40% of a cohort in honors is around 55% for quintile 2 and 35% for quintile 3, enrollment continues to rise significantly when the share of students in honors increases beyond 40%. For smaller honors programs, those in quintiles 2 and 3 that choose to enroll in honors experience an honors classroom that is likely taught to a high level and filled with high quality peers. Those that choose to not enroll in honors experience a classroom not that different from what a regular classroom would be if there were no honors classroom. Except they do not have the peer effects of the most advanced students and the teacher does not need to teach to the most advanced students.

Quintile 4 students seem to be quite insensitive with a zero effect that is only rejected for the narrow range in the share of students in honors less than 30%. The point estimate for this peak is about .03 SDs. Quintile 5 exhibits only small insignificant gains from small honors programs and once the honors program grows beyond 40%, quintile 5 students display test score losses relative to a no tracking regime. These results are consistent with the peer effect literature that has found peer effects from higher achieving students increases performance but lower achieving students are the least sensitive to 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 bottom quintile students who don't enroll in honors (which is the vast majority), the compositional changes may be offset by a better paced class. However, when the honors program grows beyond 40%, the bottom quintile students who do not enroll in honors (still the vast majority) no longer share the classroom with the middle tier students who generally share positive peer effects for the weakest tier students.

To assess the sensitivity of the results to the source of variation, I employ two alternate specifications, Figures 5a and 5b. The alternate specifications are identified by different sources of variation. The school fixed effect specification, 5a, utilizes policies that either change over time or vary by department within a school. The IV specification, 5b, utilizes across school variation and institutional momentum from policies and administrator preferences that do not change over time. The school fixed effect specification yields similar results to my primary specification, both in pattern and in magnitude. The lagged share of students in honors IV specification yields point estimates that are slightly larger in magnitude. However, this specification is noisier than the OLS specifications and is not significantly different from either my primary or the school fixed effect specifications.[24][25] If the point estimates for the IV specification are larger due to correcting for measurement error, then my primary specification will be an attenuated version of the true effect. Most importantly, my primary specification and my two alternate specifications all yield a similar pattern. Students in the top quintiles benefit significantly from honors programs with fewer than 30% of the student body in them. Students in the 2nd and 3rd quintiles benefit form honors programs with 20-40% of the student body in them. 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 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. While none of the specifications may be entirely free of endogeneity, it is comforting that the results are so consistent. For results to be severely biased, despite their consistency across specifications, one would need the biases from different sources of variation to all move in the same direction and be of similar magnitude.

My results by quintile show that honors tracking programs are not a zero sum game. Small honors programs provide a Pareto improvement by quintile with some quintiles exhibiting large gains. One can reconcile my results with papers that have found tracking doesn't harm any students if those students are primarily at schools that have small honors programs

[24]I would like to thank (Roodman, 2007) for creating the -cmp- function that allowed for this specification in Stata.

[25]The F statistics for the linear instruments are all above 390, for the quadratic instruments they are all above 290, and for the cubic instruments they are all above 150.

because only large programs harm students in bottom quintiles (Zimmer, 2003; Figlio and Page, 2002; Pischke and Manning, 2006). Similarly, one can reconcile my results with papers that have found that honors programs help top students and hurt bottom students if they sampled schools with a varied size of honors programs (Betts and Shkolnik, 2000; Hoffer, 1992; Argys et al., 1996; Epple et al., 2002).

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 allocating high quality teachers to the regular track. It is important to take precautions if a teacher has track specific human capital as reallocating them could decrease their value added performance (Cook and Mansfield, 2016).

5.2 Administrator's Problem

Recall in sect 2.2, we allow for the possibility that the optimal share of students in honors may depend on how the administrator weighs test score gains of students of different levels of preparedness. In this section I solve the administrators problem across different weighting schemes and show that the optimal size is essentially invariant to the weights chosen.[26] In particular, I consider two sets of weights, θ_qs, from equation 7, one that weighs all quintiles equally ($\theta_q = \frac{1}{5}\ \forall q$) and one that strongly prioritizes bottom quintiles so that quintiles 1, 2, 3, and 4 are weighted at 20%, 40%, 60%, and 80% of quintile 5 respectively ($\theta_q = \frac{q}{15}\ \forall q$).[27]

The left panel for Figure 6 shows the average net gains for students when each quintile is weighted the same by administrators, based on the estimates from my primary specification. Maximized gains occur for honors programs that have between 20 and 30% of students in honors, generating test score gains of 0.04 SDs relative to the absence of honors programs. The right side of Figure 6 is the average effect at the school with weights that prioritize students in bottom quintiles. The maximum weighted average effect still occurs at honors programs with enrollment shares between 20 and 30%, with a weighted average impact of 0.03 SDs. My results' robustness of the optimal honors program size across weighting schemes is driven by having gains for the top 60% of students from small honors programs and the lack of effect small honors programs have on students in the bottom 40% of preparedness.[28] Given that it's hard to shift test scores via school policies and small academic gains can have large lifetime effects, these are a sizable results.

In addition to being robust across weighting scheme, the optimal size for an honors

[26] Confidence intervals for this section are also created using the Delta Method.

[27] More weighting schemes are available upon request from the author.

[28] When using point estimates for treatment effects and having every quintile be given a weight of at least 0.1, smaller honors programs dominate larger ones. Increasing the share of students in honors beyond 35% decreases gains for every weighting scheme for the remaining support of the data.

program is robust across specifications. Figure 7 displays the average effect for my alternate specifications under both weighting schemes. The school fixed effect specification, on the left side of Figure 7, has a smaller weighted average, but the optimal share in honors remains between 20 and 30%. The IV specification has larger point estimates for the weighted average gains 7, but the same optimal share of honors. This could be due to increased noise, increased bias, or correcting for measurement error.

These effects combine the impacts of several mechanisms, including specialized instruction, allocation of teachers, and peer effects. In order to decompose these effects, stronger additional assumptions would be required. Note though that identifying the effect of changing the size of honors programs, a key policy lever for administrators, does not require separate identification of these mechanisms.

My data do not included career and later life outcomes. However, one can perform a back of the envelope calculation of the effect of estimated test score gains on lifetime outcomes by assuming that test score gains from varying the size of honors programs has the same effect as test score gains from teacher quality, found in (Chetty et al., 2014b,a)[29]. Students at schools with small honors programs would have their earnings at age 28 increase by an average of 0.4% compared to if their school had no honors programs for each core course in my sample. So if a high school with a class of 100 students switched from a scheme with no honors program to an optimal honors program for all six classes in my sample, those students would have their expected aggregate earnings at the age of 28 increase by over $88,000.[30] If other courses not tested, such as English classes other than English 1, had similar effects then this estimate could be much larger. If all schools in my sample switched from their current honors program size to an honors program with 20 to 30% of the student body in it, the average student would create a test score gain of over 0.02 SDs (about the same amount as switching from the median teacher to a 57^{th} percentile teacher (Chetty et al., 2014a)). North Carolina averages about 100,000 students per grade per year. If all the high schools in North Carolina switched to the optimal honors program size, then the aggregate increase in earnings for 28 yearolds state wide would be over $44 million. While this exercise is quite speculative, it highlights the possiblity that small student gains from a superior tracking system can none the less aggregate to very large earnings contributions when considering the effect across all courses and all years.

[29] This is plausible as some of the mechanisms through which honors track size affects test scores are peer effects, specialized instruction, and teacher assignment. The former of these has also been shown to have large effects on lifetime earnings (Carrell et al., 2018).

[30] This assumed all students have a median income of $36910 at the age of 28. This level was chosen based on the 2018 median income for 28 year-olds.

6 Robustness Checks

To ensure that the restriction that having all students in honors is the same as having no students in honors is not driving my results I run an unrestricted version of my primary specification, Appendix Figure A.7a. The effect of changing the size of an honors track is similar to my restricted version over the support of my data as the restriction is only meaningful when extrapolating to points outside of the support that no administrator is considering. To help visualize this Appendix Figure A.7b has my primary and the unrestricted specifications on the same graphs. This also indicates that there is little reason to believe that there is a large curricula effect, large rigor effect, or other effect not driven by the size of honors. As a result, I am comfortable with my analysis assuming that having all classes be honors classes has the same effect as having none of the classes be honors classes.

By not including an indicator for the existence of an honors program I have increased the precision of my main results at the cost of having my specification rule out the possibility of a discontinuity at zero. Appendix Figure A.8a is the same as my primary specification, but an indicator for whether an honors program exists is introduced. The results are similar to my primary specification. However, by introducing the indicator for the existence of honors and interacting it with quintiles, my precision drops significantly for shares of students in honors with limited observations. To show this Appendix Figure A.8b combines my primary specification with this this one.

I employ a specification that uses the share of classrooms that are honors as an instrument for the share of students in honors. Administrators generally set the offered classes then students select into either the honors or regular track. If administrators set the track offerings based on observable cohort characteristics, then this specification not be biased by the unobserved quality of a cohort. Appendix Figure A.9 in Appendix A is the treatment effect when the fraction of students in honors is determined by a first stage utilizing the fraction of classrooms that are honors. This specification yields results that look similar to my primary specification. Hence, when aggregating to average effects, this IV specification also matches my primary specification as seen by Figure A.10.

Another check to see if my results are being driven by unobservably better or worse cohorts is through school-year fixed effects. The limitation to school-year fixed effects is they are purged of all variation in the share of students in honors except within year-across-course variation. Appendix Figure A.11 are the results to my primary specification with added school year fixed effects. The pattern is the same for quintiles, however the effect sizes are very muted. This could be due to purging most of the variation in the share of students in honors. The muted effect sizes will prevent the aggregate treatment effect from

being significantly different from zero. Hence, I do not include the aggregate version in this paper.[31]

It is possible that the subset of school-courses used in section 4 is not restrictive enough for Assumption 2 from Section 2.2 to be valid. To test whether my set of schools includes observations that violate Assumption 2, my primary specification is reran on a schools that where shifting each student less than a third of a quintile on average could achieve a uniform distribution. Figure A.12 in Appendix A is the same as my primary specification, but with the restricted sample. The noise has increased, but the point estimates are roughly the same. As a result, I am confident that the cutoff used for section 4 does not invalidate Assumption 2.

7 Conclusion

In this paper I use very rich administrative data to identify the effects of alternative honors enrollment shares separately by level of student preparedness, allowing for endogenous self-sorting of students into the honors track conditional on the existing share. The estimates are a sufficient set of inputs to solve the administrator's problem of what fraction of the students should be in the honors track. I recover the very robust result that the optimal share of students in the honors is between 20 and 30%. Using the results of my primary specification, if all the schools switched from their current honors program sizes (including the absence of an honors program) to one with 20 to 30% of students in it, students in my sample would gain over 0.02 SDs in test score performance on average. These gains are not at the expense of harming students of a certain ability, as no quintile is hurt by small honors programs. These results are robust to alternative specifications that utilize different sources of variation and remove different sources of endogeneity, making it more likely that my results capture the true effects

My results can guide administrators at high schools where students have a say in their tracking decision to a low cost method to improve test score performance. If parents are uncertain of the child's unobserved ability or preferences towards certain tracks, this could also provide them with information on what schools have tracking policies that will likely benefit their child. To give parents information on whether they should enroll in honors or not requires estimates of a different parameter that other regression discontinuity approaches and propensity score matching papers have done.

Although my final sample did exclude many schools that had a disproportionate number of advanced or struggling students, the external validity of my results is great. Programs

[31] These results are available per request from the author.

where different tracks have different curricula, such as AP or IB, may experience similar gains from a small advanced track. Unfortunately, it will be difficult to test the effectiveness of tracking a relatively small share of students into the more advanced track for these programs as they test different curricula. Further assumptions will be required to convert test score performance in different tests to a educational numeraire. For programs outside of the state and non-honors programs for high school students, the external validity likely applies to systems that gives students and/or parents some choice into track selection.

References

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.

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.

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. Statistical Software Components, Boston College Department of Economics, 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.

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

Tables

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

Table 1: Frequency of tracking offerings by course. Observations are limited to cohorts with at least 30 observations.

VARIABLES	No honors tracking mean (sd)	Share $\in (0, 0.35)$ mean (sd)	Share $\in [0.35, 1)$ mean (sd)
Pupil-teacher ratio	15.34	15.63	16.06
	(2.775)	(2.676)	(2.945)
Title 1 status	0.999	0.992	0.993
	(0.0382)	(0.0908)	(0.0857)
School performance	74.47	73.53	74.08
	(8.240)	(8.260)	(7.977)
Cohort size	97.85	266.3	304.5
	(43.85)	(154.7)	(271.7)
Average Praxis scores	523.2	527.9	515.1
	(194.6)	(154.5)	(174.7)
Teacher share with			
Bachelor's	0.912	0.887	0.874
	(0.236)	(0.229)	(0.250)
Master's	0.288	0.257	0.282
	(0.384)	(0.325)	(0.346)
Advanced degree	0.00601	0.0105	0.00918
	(0.0644)	(0.0706)	(0.0705)
Doctorate	0.0136	0.00382	0.00414
	(0.103)	(0.0469)	(0.0528)
Standard professional II licenses	0.917	0.922	0.911
	(0.238)	(0.189)	(0.213)
Standard professional I licenses	0.0531	0.0509	0.0603
	(0.194)	(0.154)	(0.177)
Provisional licenses	0.0162	0.0132	0.0119
	(0.107)	(0.0860)	(0.0799)
Temporary licenses	0.0167	0.0179	0.0202
	(0.100)	(0.0891)	(0.0968)
0 years exp	0.0680	0.0501	0.0628
	(0.214)	(0.165)	(0.190)
1 year exp	0.0284	0.0289	0.0330
	(0.139)	(0.120)	(0.129)
2 years exp	0.0425	0.0381	0.0340
	(0.178)	(0.137)	(0.133)
3-5 years exp	0.0883	0.0948	0.0895
	(0.245)	(0.216)	(0.222)
6-11 years exp	0.237	0.218	0.226
	(370)	(0.308)	(0.325)
12+ years exp	0.536	0.570	0.555
	(0.438)	(0.373)	(0.393)
Fraction of students			
White	0.691	0.690	0.705
	(0.228)	(0.201)	(0.187)
Black	0.241	0.232	0.224
	(0.217)	(0.184)	(0.173)
Hispanic	0.0485	0.0514	0.0498
	(0.0434)	(0.0427)	(0.0432)
Migrant	0.00772	0.00987	0.00399
	(0.0283)	(0.0287)	(0.0163)
With gifted status	0.165	0.134	0.187
	(0.122)	(0.0989)	(0.133)
With diagnosed learning disabilities	0.0128	0.0388	0.0292
	(0.0175)	(0.0301)	(0.0267)
With free lunch	0.264	0.271	0.248
	(0.0952)	(0.109)	(0.101)
With reduced lunch	0.0767	0.0740	0.0638
	(0.0385)	(0.0365)	(0.0361)
With free or reduced lunch	0.341	0.345	0.312
	(0.114)	(0.125)	(0.119)
Whose parents lack a HS diploma	0.0682	0.0658	0.0451
	(0.0456)	(0.0388)	(0.0318)
Whose parents have a HS diploma	0.244	0.184	
	(0.0797)	(0.0724)	(0.0742)
Whose parents have some college	0.144	0.141	0.136
	(0.0496)	(0.0368)	(0.0453)
Whose parents attended trade or business school	0.0366	0.0331	0.0350
	(0.0235)	(0.0177)	(0.0206)
Whose parents attended community college	0.208	0.199	0.183
	(0.0625)	(0.0527)	(0.0660)
Whose parents have a 4-year degree	0.210	0.227	0.280
	(0.0776)	(0.0755)	(0.0839)
Whose parents have graduate degrees	0.0802	0.0820	0.128
	(0.0579)	(0.0484)	(0.0783)
School-course-years	687	2,286	1,759

Table 2: All statistics are school-course-year averages. Sample is the same as my primary specification. Observations are limited to cohorts with at least 30 observations. Schools with atypical distributions of student quality have been dropped.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quintile 1							
Linear Coefficient	0.727***	0.549***	0.981***	0.672***	0.700***	0.826***	0.501***
	(0.143)	(0.114)	(0.197)	(0.154)	(0.147)	(0.166)	(0.114)
Quintile 1							
Squared Coefficient	-1.957***	-1.517***	-2.754***	-1.830***	-1.945***	-2.231***	-1.452***
	(0.414)	(0.308)	(0.569)	(0.438)	(0.432)	(0.464)	(0.312)
Quintile 1							
Cubic Coefficient	1.230***	0.968***	1.773***	1.097***	1.245***	1.405***	0.952***
	(0.284)	(0.206)	(0.388)	(0.316)	(0.299)	(0.312)	(0.211)
Quintile 2							
Linear Coefficient	0.438***	0.251***	0.629***	0.442***	0.391***	0.516***	0.210**
	(0.124)	(0.0949)	(0.169)	(0.130)	(0.131)	(0.136)	(0.0971)
Quintile 2							
Squared Coefficient	-1.053***	-0.602**	-1.698***	-1.148***	-0.904**	-1.283***	-0.531**
	(0.360)	(0.257)	(0.494)	(0.378)	(0.382)	(0.390)	(0.267)
Quintile 2							
Cubic Coefficient	0.615**	0.350**	1.070***	0.725***	0.513*	0.768***	0.321*
	(0.249)	(0.173)	(0.340)	(0.273)	(0.265)	(0.267)	(0.183)
Quintile 3							
Linear Coefficient	0.274**	0.131	0.532***	0.316**	0.246*	0.325***	0.0701
	(0.120)	(0.0953)	(0.164)	(0.126)	(0.127)	(0.134)	(0.102)
Quintile 3							
Squared Coefficient	-0.526	-0.194	-1.349***	-0.708*	-0.455	-0.668*	-0.101
	(0.347)	(0.256)	(0.484)	(0.378)	(0.370)	(0.390)	(0.277)
Quintile 3							
Cubic Coefficient	0.252	0.0628	0.817***	0.425	0.209	0.343	0.310
	(0.239)	(0.172)	(0.335)	(0.284)	(0.209)	(0.268)	(0.187)
Quintile 4							
Linear Coefficient	0.280**	0.129	0.426***	0.371***	0.226*	0.387***	0.0561
	(0.117)	(0.0918)	(0.159)	(0.127)	(0.125)	(0.133)	(0.0967)
Quintile 4							
Squared Coefficient	-0.722**	-0.385	-1.188**	-1.085***	-0.56 2	-1.025***	-0.260
	(0.339)	(0.259)	(0.480)	(0.374)	(0.363)	(0.384)	(0.271)
Quintile 4							
Cubic Coefficient	0.443*	0.257	0.762**	0.791***	0.336	0.638**	0.204
	(0.234)	(0.178)	(0.335)	(0.273)	(0.250)	(0.265)	(0.187)
Quintile 5							
Linear Coefficient	0.204*	0.0594	0.346**	0.305**	0.186	0.312***	-0.0281
	(0.107)	(0.0971)	(0.154)	(0.119)	(0.116)	(0.120)	(0.101)
Quintile 5							
Squared Coefficient	-0.706**	-0.387	-1.234***	-1.133***	-0.669**	-1.048***	-0.192
	(0.313)	(0.284)	(0.456)	(0.354)	(0.340)	(0.354)	(0.293)
Quintile 5							
Cubic Coefficient	0.502**	0.328*	0.887***	0.912***	0.483**	0.736***	0.220
	(0.217)	(0.199)	(0.315)	(0.263)	(0.235)	(0.247)	(0.205)
Observations	118,866	118,866	110,377	118,866	118,866	90,806	118,866
School FEs	NO	YES	NO	NO	NO	NO	NO
School-year FE	NO	NO	NO	NO	NO	NO	YES
Constrained Coefficients	YES	YES	YES	NO	YES	YES	YES
Lagged IV	NO	NO	YES	NO	NO	NO	NO
Classroom Share IV	NO	NO	NO	NO	YES	NO	NO
Restricted Sample	NO	NO	NO	NO	NO	YES	NO

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3: All standard errors are clustered at the school level.

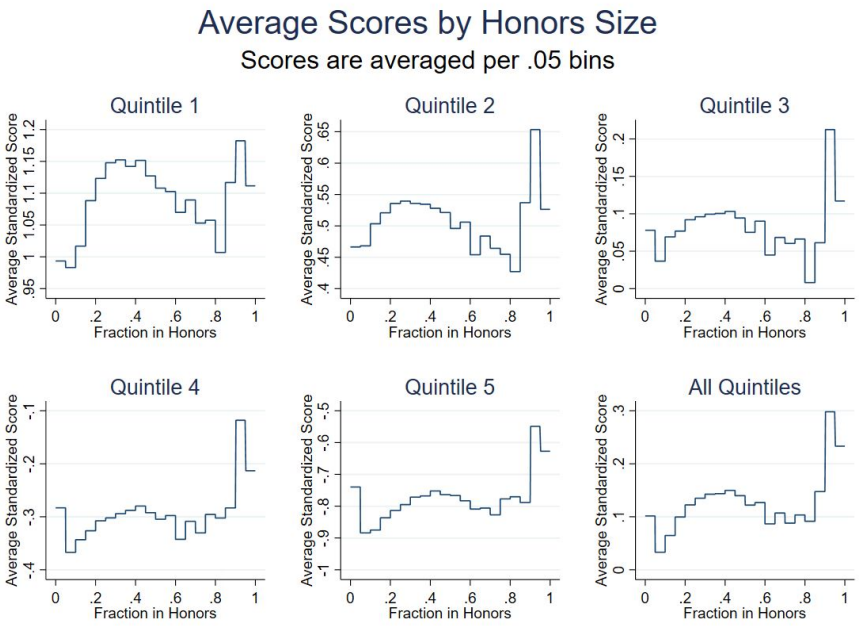


Figure 1: The bin for the lowest share of students in honors include school-year-courses where none of the students are in honors. For the rest of the bins, they include shares in (bin minimum, bin maximum]. No schools with IB are included. All school-year-courses have at least 30 observations.

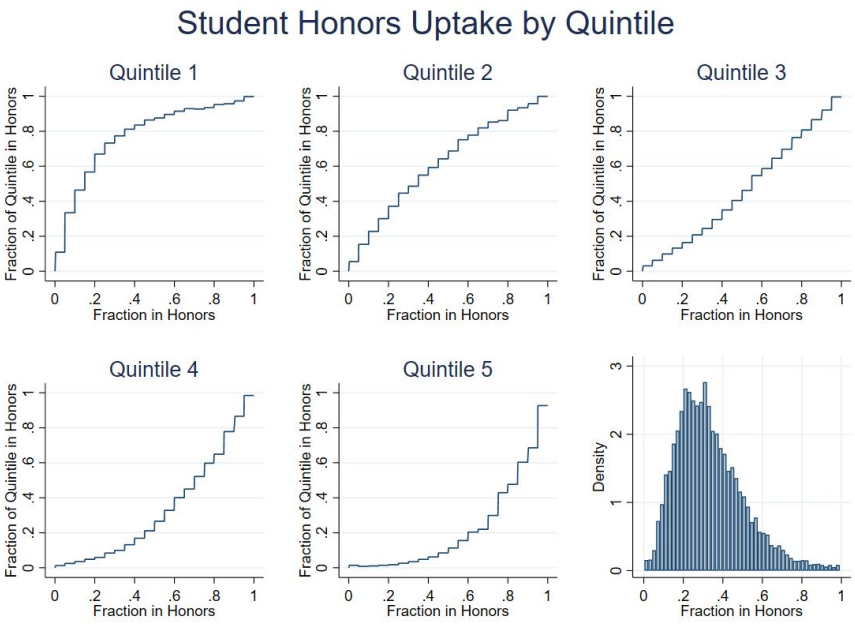


Figure 2: For this figure, the quintiles are within school predicted performance rankings. Each bin includes shares in (bin minimum, bin maximum]. The sixth cell shows the support of the data, excluding school-year-courses where either none of the students or all of the students are enrolled in honors. No schools with IB are included. All school-year-courses have at least 30 observations.

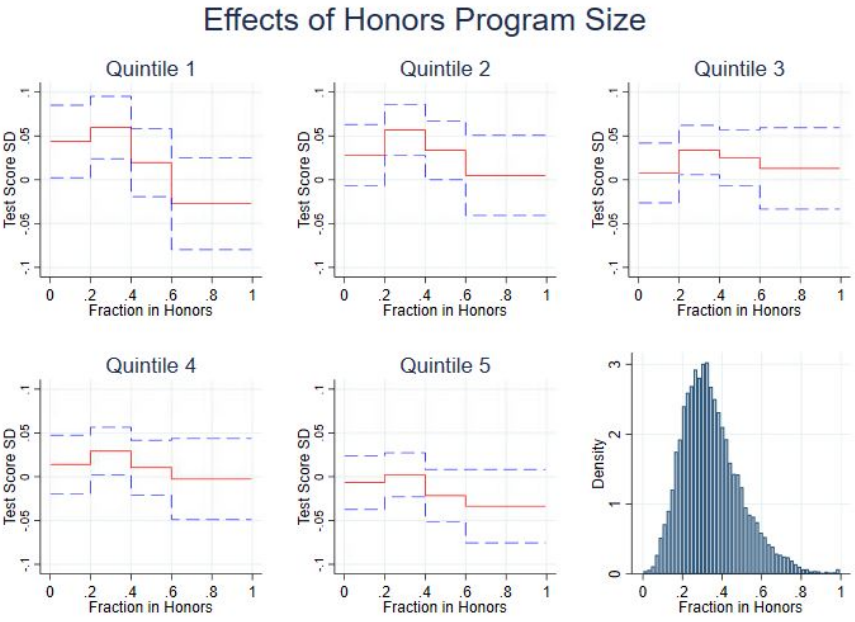


Figure 3: This specification has five indicators for the share of students in honors. The indicator for having no students in honors is excluded. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals are shown.

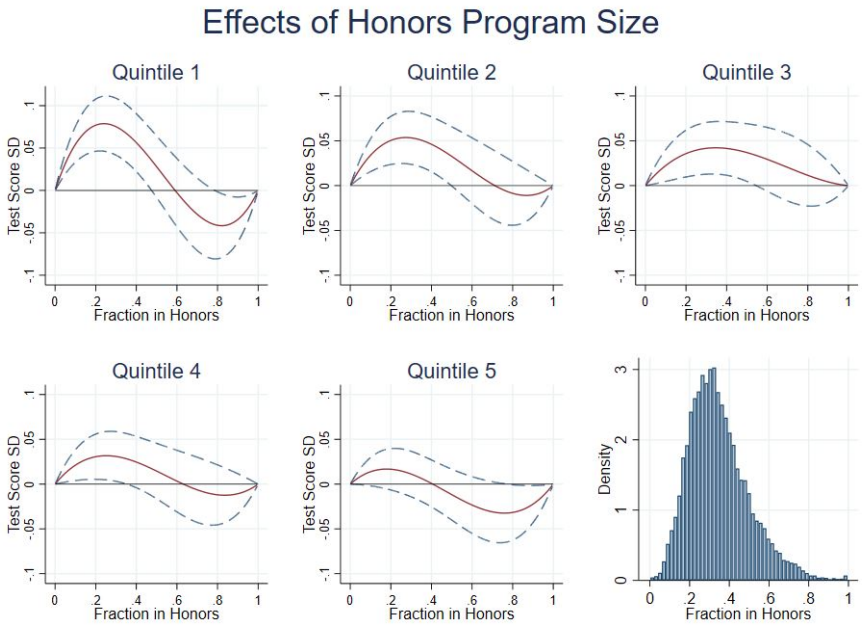
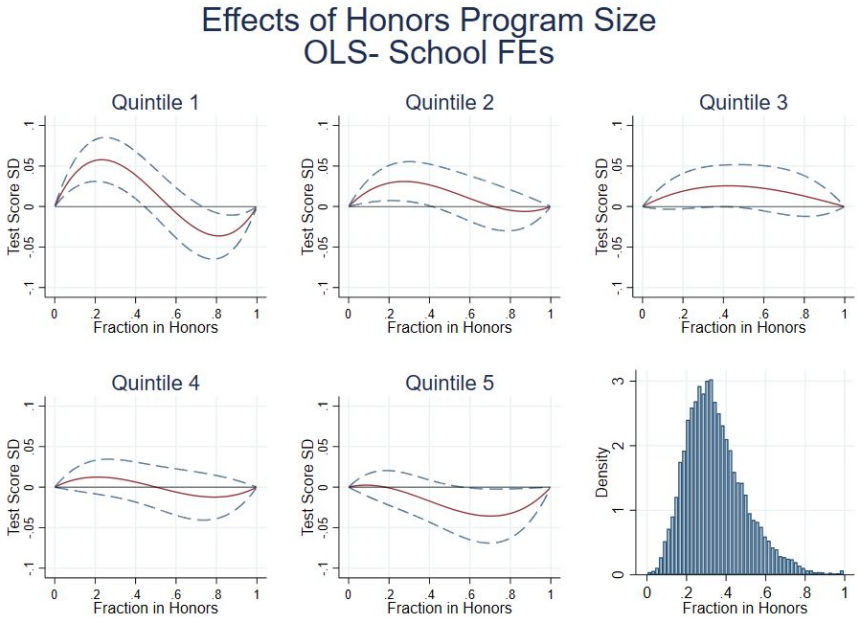
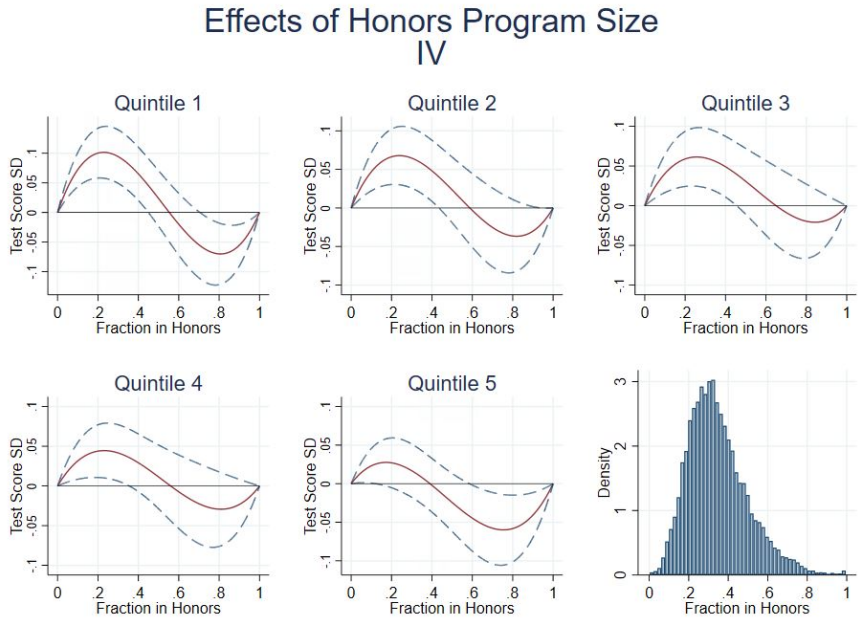


Figure 4: Sample is limited to schools where shifting students on average $\frac{1}{2}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.



(a)



(b)

Figure 5: For the IV specification the current share of honors is instrumented with the previous share of honors at each school-course. Sample is limited to schools where shifting students on average $\frac{1}{2}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.

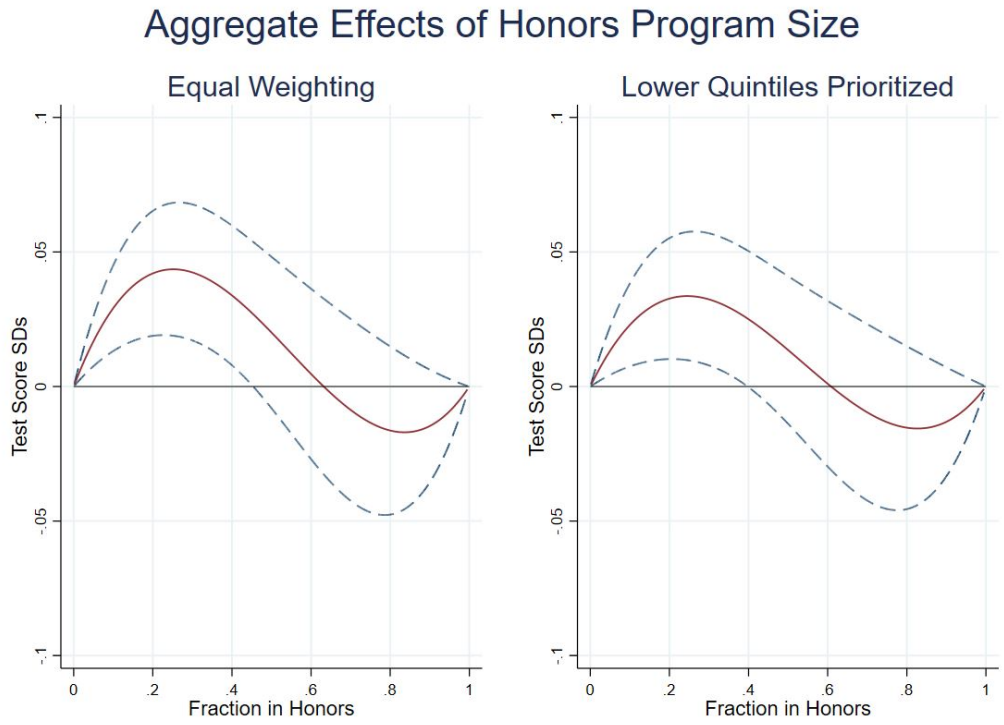


Figure 6: For the specification that prioritizes bottom quintiles, the weighting scheme assigns the following weights to observations in quintiles 1, 2, 3, 4, and 5: $\frac{1}{15}$, $\frac{2}{15}$, $\frac{3}{15}$, $\frac{4}{15}$, and $\frac{5}{15}$. Sample is limited to schools where shifting students on average $\frac{1}{2}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.

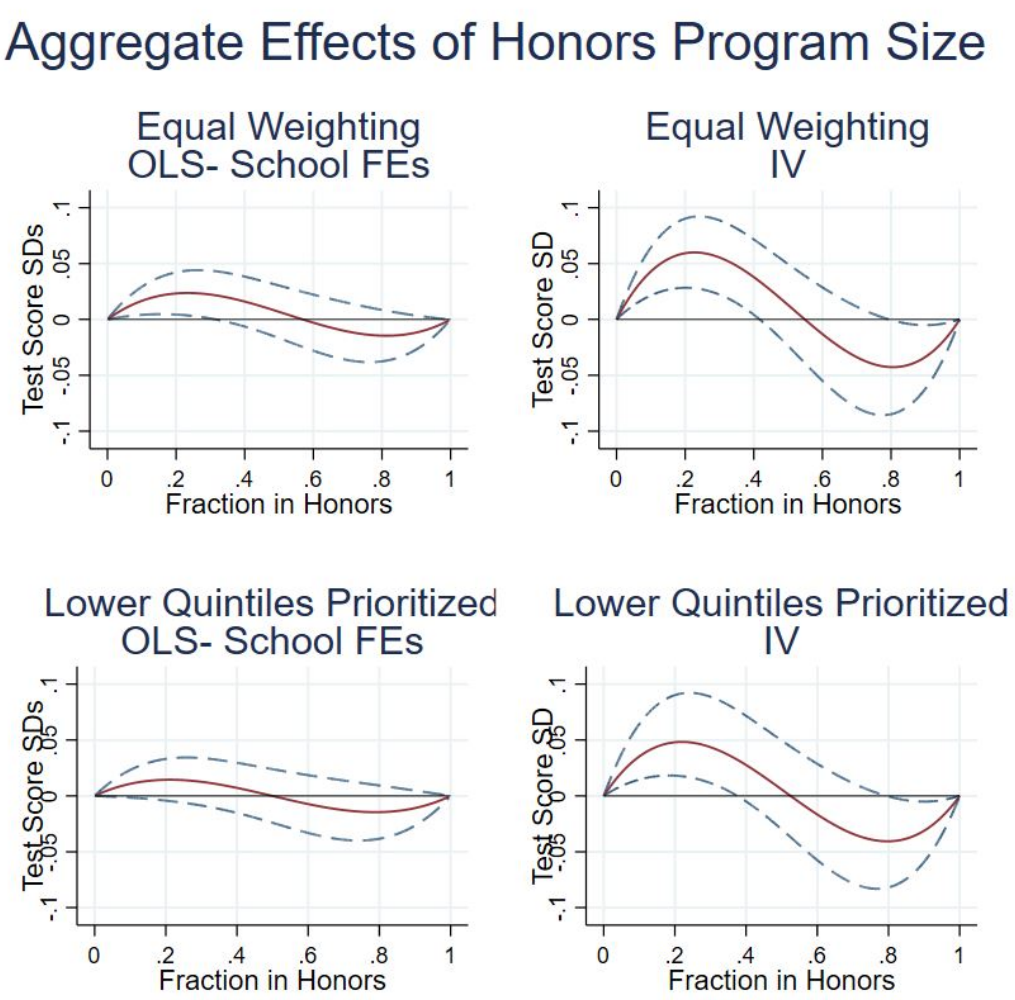


Figure 7: For the specification that priortizes bottom quintiles, the weighting scheme assigns the following weights to observations in quintiles 1, 2, 3, 4, and 5: $\frac{1}{15}$, $\frac{2}{15}$, $\frac{3}{15}$, $\frac{4}{15}$, and $\frac{5}{15}$. Sample is limited to schools where shifting students on average $\frac{1}{2}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.

A Appendix

Figures

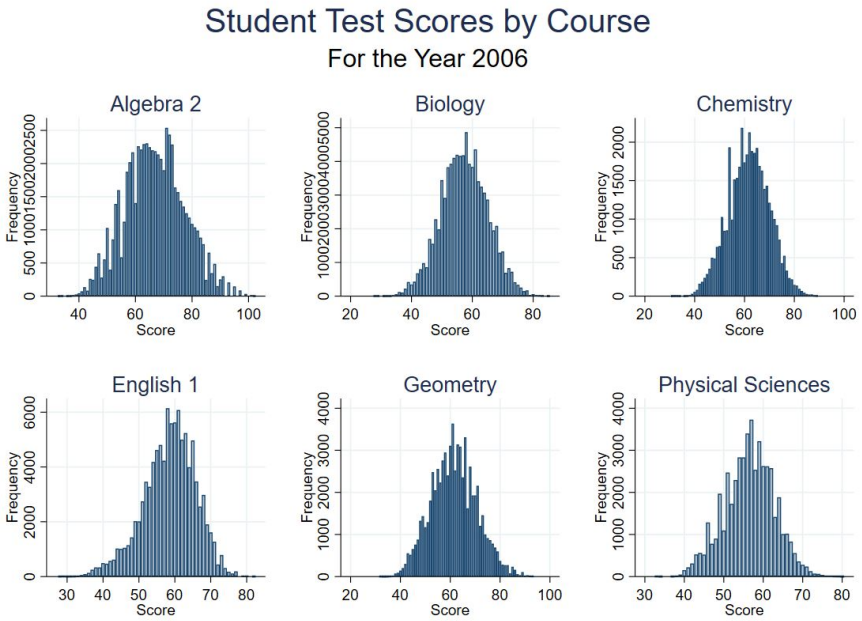


Figure A.1: The score for student standardized test performance in the year 2006, for each course included in my final sample. The figure indicates that there is not grouping near the ceiling nor the floor of the test score range. More years are available by request.

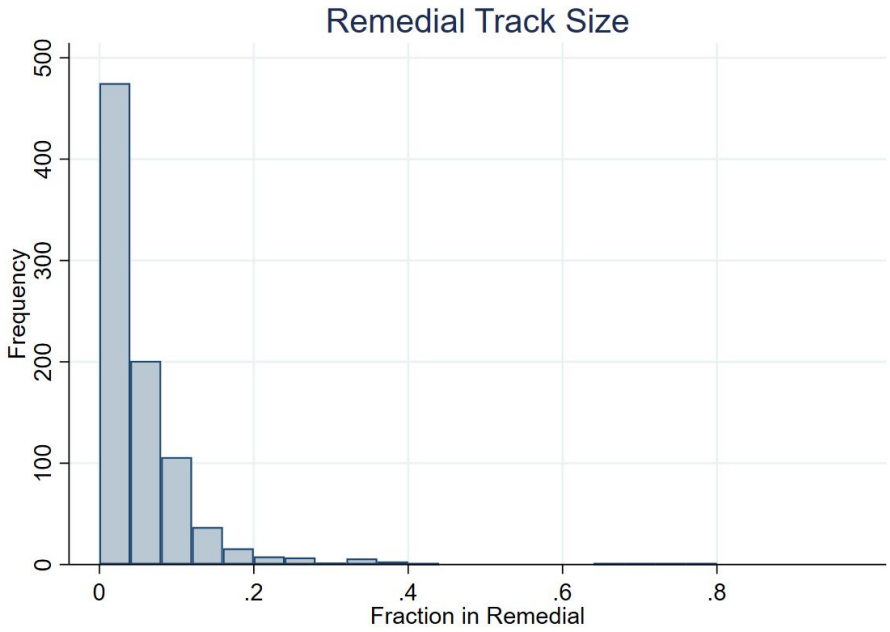


Figure A.2: This figure shows the fraction of students in the remedial track for school-year-courses where there is a remedial track. Fewer than 4% of school-year-courses in my sample have a remedial track. No schools with IB are included. All school-year-courses have at least 30 observations.

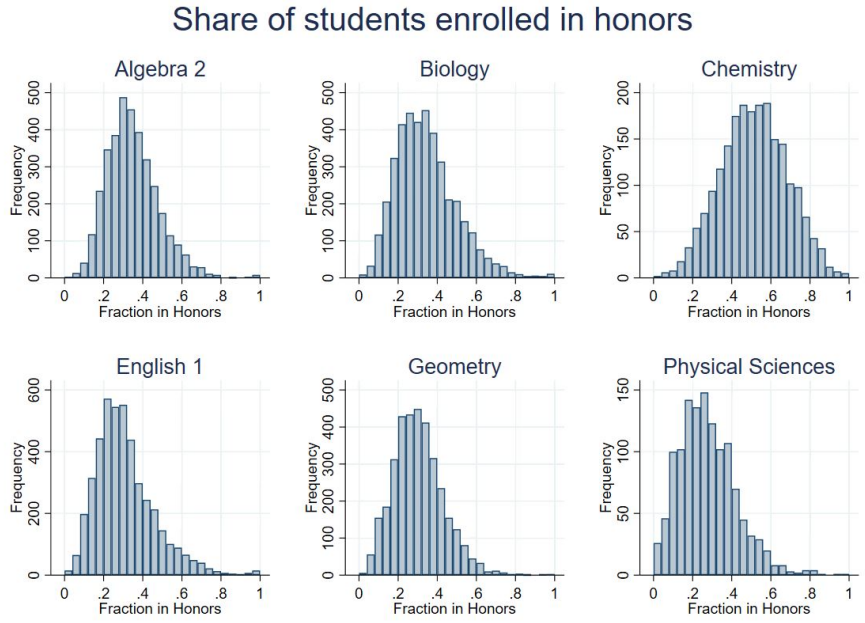


Figure A.3: This figure shows the distribution of the fraction of students in the honors track for school-year-courses where there is an honors track. No schools with IB are included. All school-year-courses have at least 30 observations.

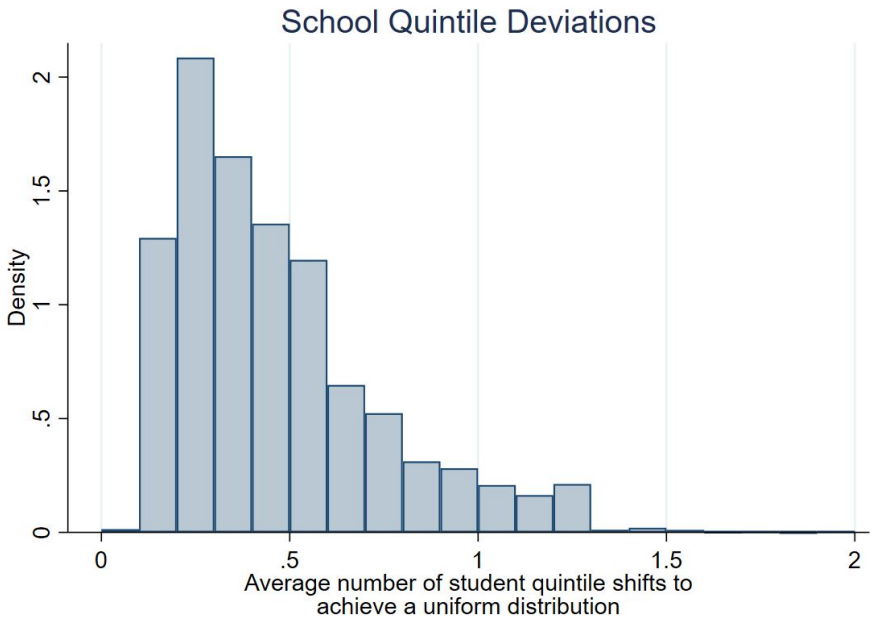


Figure A.4: This figure shows the distribution of the how many quintiles the students would have to shift on average to have a uniform distribution of across quintiles of student preparedness at schools in my sample.

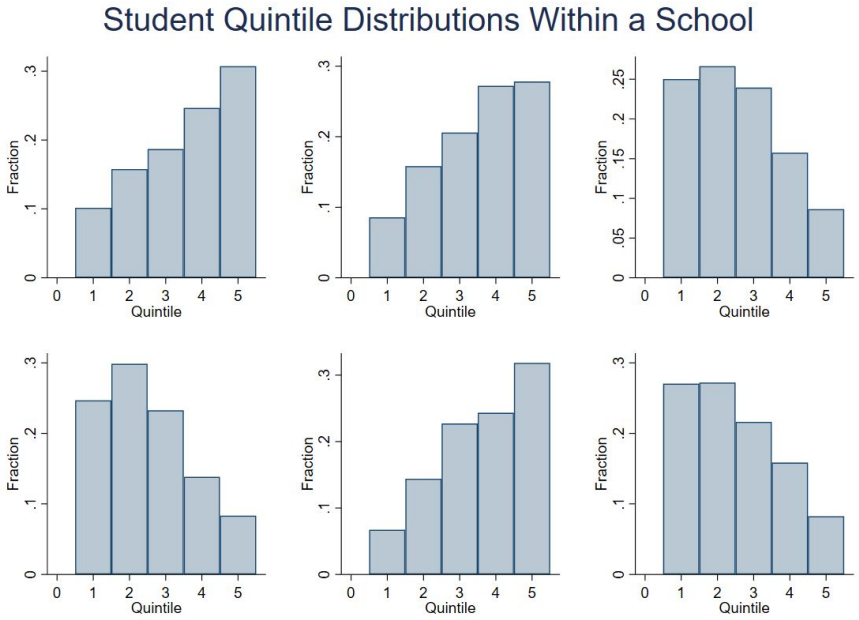


Figure A.5: The 6 schools with observed quality distributions such that shifting each student an average of half a quintile will yield a uniform distribution. These distributions are the least uniform distributions included in my primary specification.

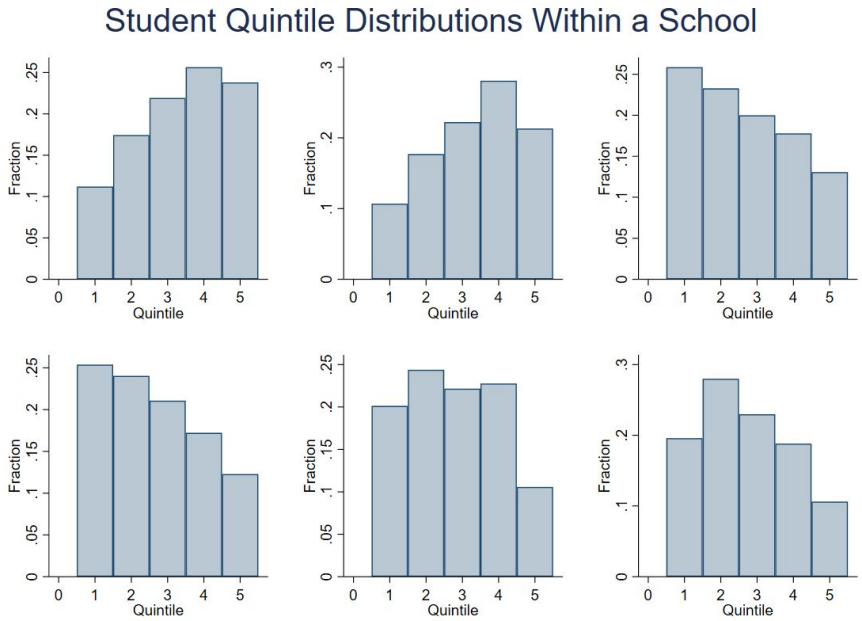
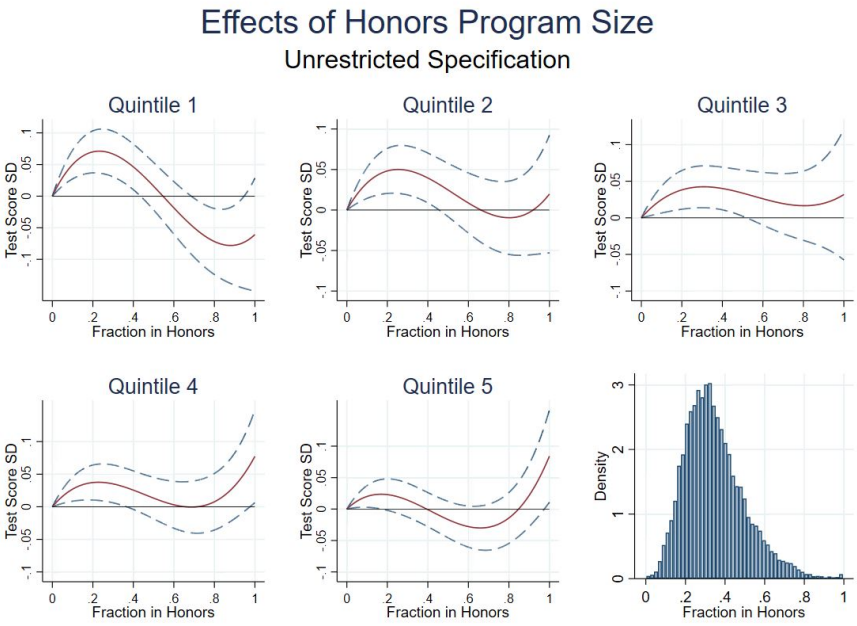
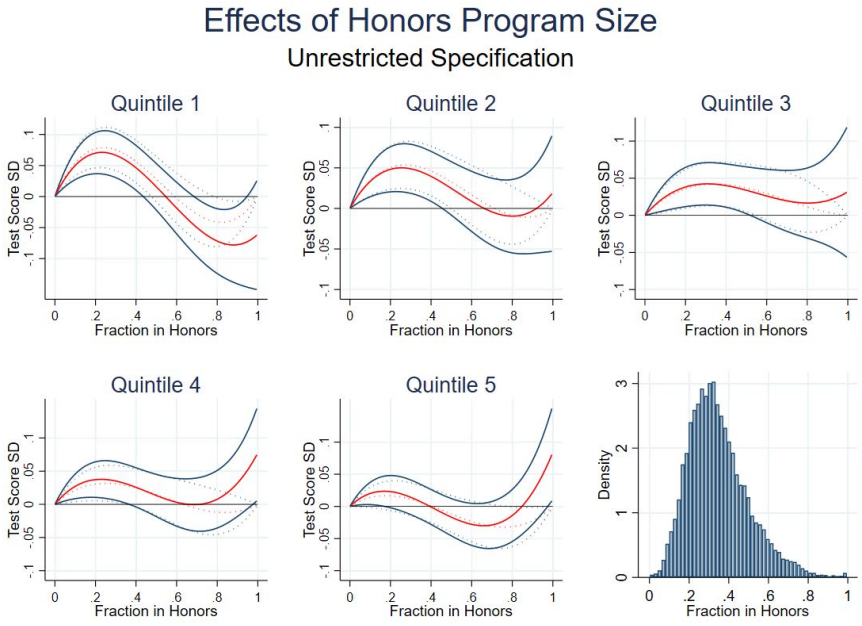


Figure A.6: The 6 school-courses with observed quality distributions such that shifting each student an average of a third a quintile will yield a uniform distribution

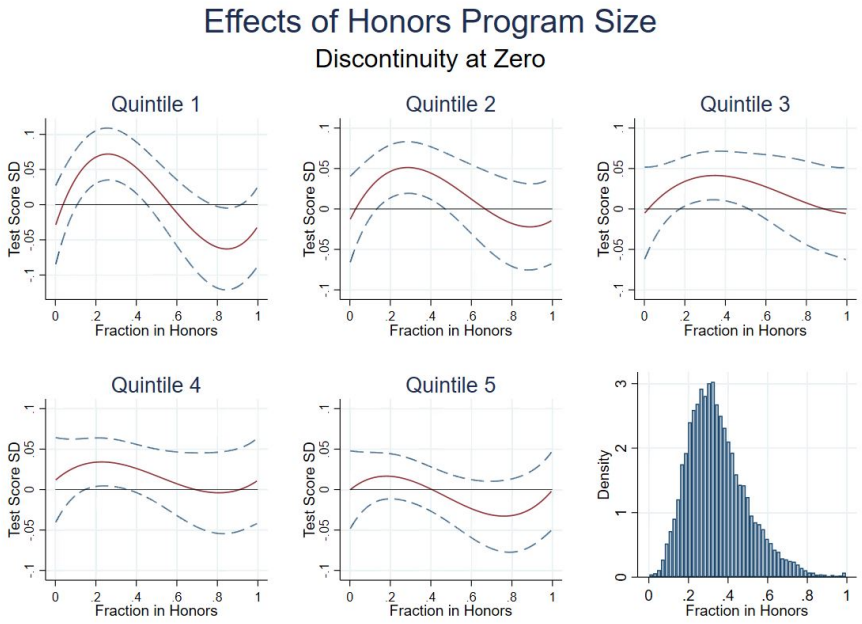


(a)

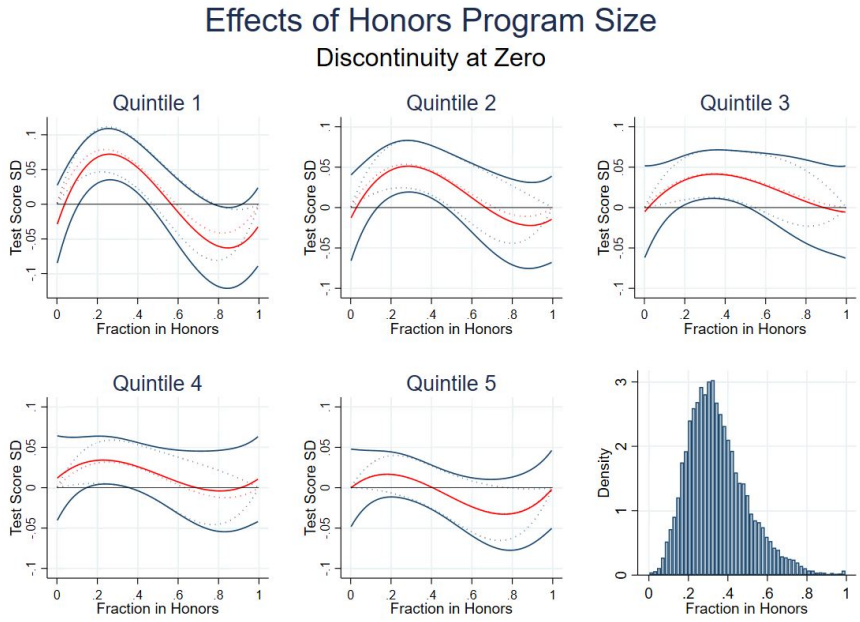


(b)

Figure A.7: Figure (a) is a version of my primary specification that does not restrict the treatment effect to be the same when all students and no students are in honors. Figure (b) has the results from figure (a) presented with the solid lines and the results from my main specification presented with dotted lines. Sample is limited to schools where shifting students on average $\frac{1}{2}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.



(a)



(b)

Figure A.8: Figure (a) is a version of my primary specification with an additional indicator for whether there is an honors program. Figure (b) has the results from figure (a) presented with the solid lines and the results from my main specification presented with dotted lines. Sample is limited to schools where shifting students on average $\frac{1}{2}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.

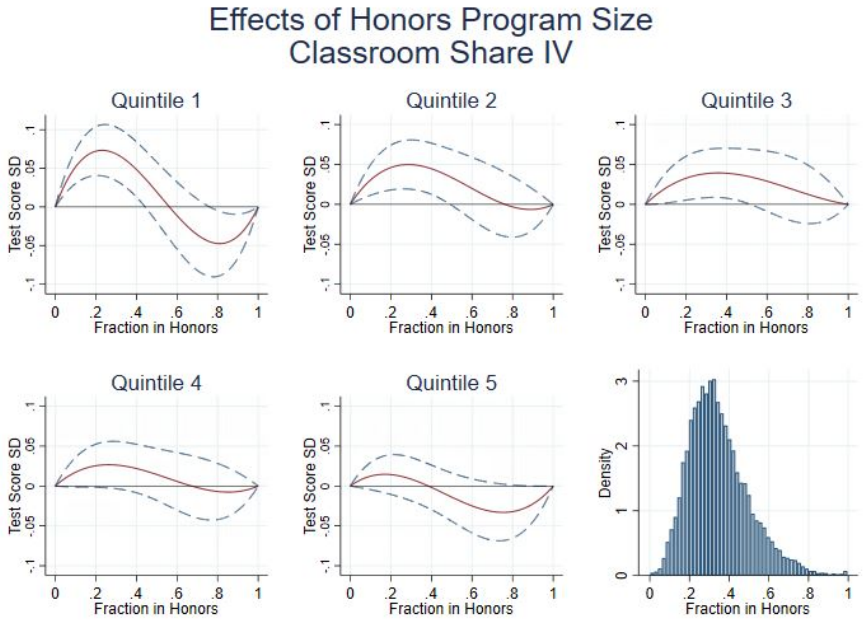


Figure A.9: The school-year-course fraction of students in honors is instrumented with the school-year-course fraction of classes that are honors. Sample is limited to schools where shifting students on average $\frac{1}{2}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.

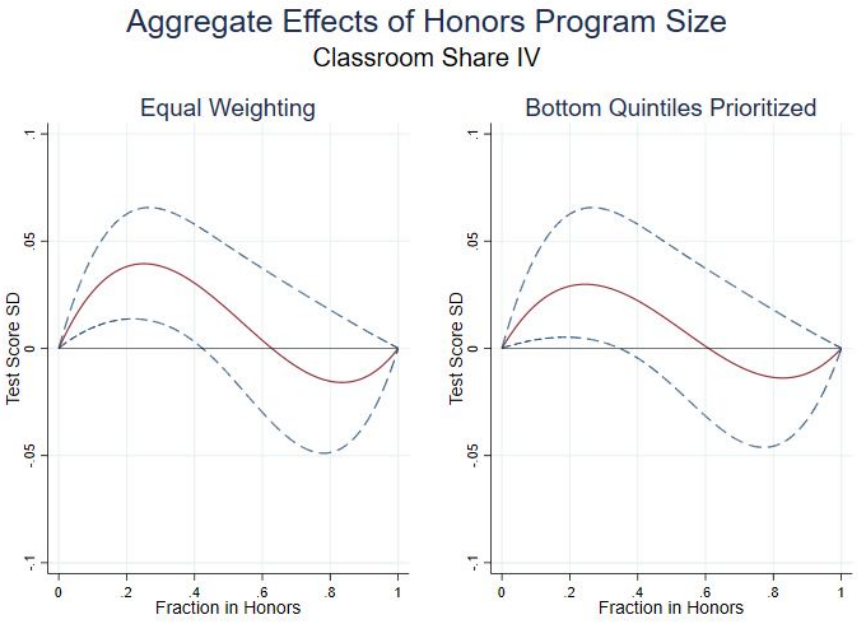


Figure A.10: For the specification that priortizes bottom quintiles, the weighting scheme assigns the following weights to observations in quintiles 1, 2, 3, 4, and 5: $\frac{1}{15}$, $\frac{2}{15}$, $\frac{3}{15}$, $\frac{4}{15}$, and $\frac{5}{15}$. The school-year-course fraction of students in honors is instrumented with the school-year-course fraction of classes that are honors. Sample is limited to schools where shifting students on average $\frac{1}{2}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.

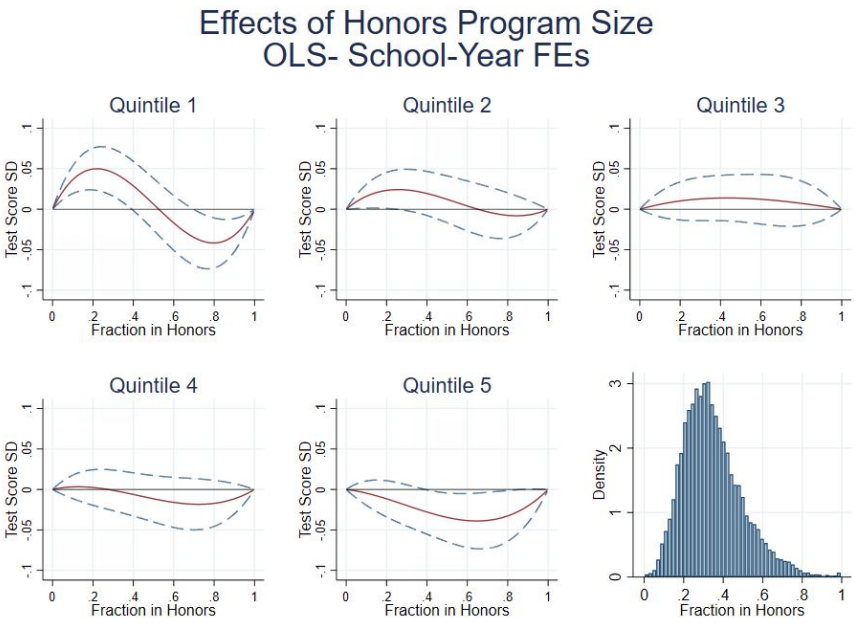


Figure A.11: Sample is limited to school-courses where shifting students on average $\frac{1}{2}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.

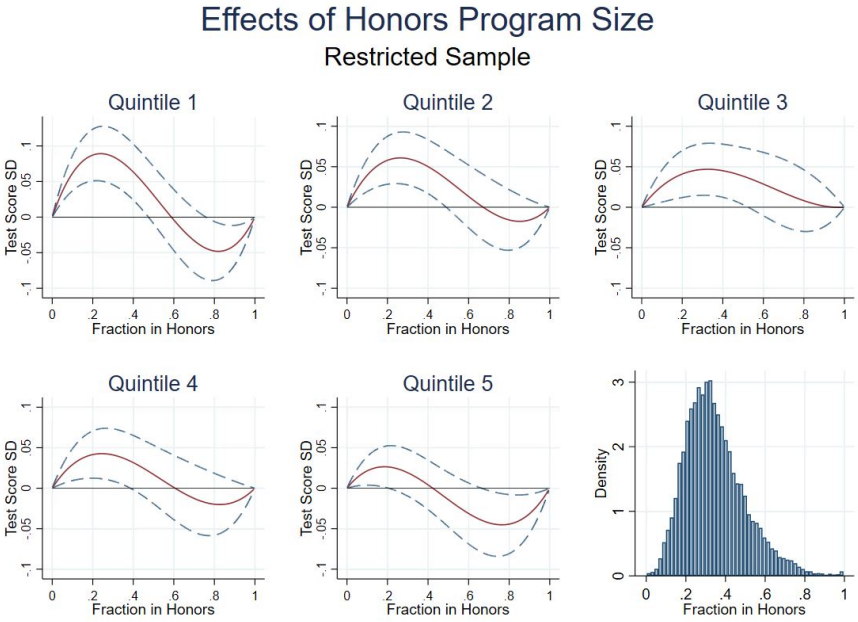


Figure A.12: Sample is limited to school-courses where shifting students on average $\frac{1}{3}$ a quintile or less can achieve a uniform distribution. No schools with IB are included. All school-year-courses have at least 30 observations. 95% confidence intervals shown.