1 Informal Summary Of Results Across Specifications

- Quintile and Quintile 2 have a positive, significant peak between .2 and .4 in essentially every specification.
- Quintile 3 frequently shows a slight, non-significant gain from small honors programs, and never shows meaningful losses from small honors programs (or really any honors program).
- Quintile 4 seems to be essentially unaffected by small honors programs (some specs positive, some specs negative), and occasionally shows significant losses at high honors shares.
- Quintile 5 is also negligibly affected by very small honors programs, but begins to reveal losses around .2 to .4 and becomes significant after .4.

Items still to do

1. Are we dropping classes with fewer than, say, 5 students? It seems like we have a surprisingly large number of classrooms per course!
2. It would be nice to plot the distribution of number of classrooms per course across all courses (or the mean number by school).
3. Seems like we may have posted the Class share IV results for the administrator problem instead of the other course IV. Can we check this?
4. Table with administrator problem predicted values.
5. Should run the model on the full sample of schools (ignoring the 0.5 per-student quintile shift restriction).
6. Discuss how we produce the simulated gains from reassigning honors shares.
7. Add several controls. Share free/reduced price lunch, ever free reduced price lunch. English as a second language. District controls
8. How well can share honors across school-years be predicted based solely on the predicted cohort-level quintile distribution? Can then re-do for school-course-year level or both left and right hand side. - Zach/Rick

9. How much variation in honors boosts do you observe in the data? I.e. what share of schools and school-years report this, and what are the most frequently used boosts - Zach

10. Need a more extensive Appendix summary statistics table with all of the controls: all baseline controls, separated by the level at which they vary (school-level vs. stj vs. stjq). Report mean, standard deviation, and min/max for each variable. - Zach

11. May need a more detailed glossary for each variable. (Could be a long footnote to the above table) - Zach.

12. Mention the fact that AP is usually taken in the second course for Chemistry and Biology and English, Physics sometimes lets first year take AP, so we drop. U.S. history only has 1 course, so we drop. No AP geometry or Algebra 1 or 2. Rick

13. Is the requirement that all school-year-course have 30 observations imposed on all samples? YES, IT IS. If so, we don't need to say it in each figure not, but should make this clear in the paper. Also, are these 30 observations test scores or students? Test scores! - Rick

14. Have Zach read the paper and make suggestions.

15. Should run a specification where we remove controls for class size at the course level, since changing the share honors might mechanically change class size, so this might be construed as part of the effect.

16. What years are used? Because we need lags of test scores, and some courses stopped providing standardized testing. 1999-2011. Middle school data comes from earlier.

17. Compute the number of test scores , students, classrooms and schools (124,524 school-course-year-deciles, 1058 schools ever, 716) in the sample. Put in Section 3.2. Are confidence intervals 90% everywhere or 95% intervals in some places?

18. Do we need to update the Chetty teacher quality percentiles and predicted lifetime earnings gains?

2 Other Notes

1. Specification choices: which are main specifications, which are robustness checks, which don't enter the paper at all.

2. Possible specifications: Baseline, all variation
   Threats to validity:

3. unobservably superior distribution of students causes a larger share of students in honors. Could occur at both the school, school-cohort, or even school-cohort-course level.

4. Unobservably superior distribution of administrators causes a larger share of students in honors. Mostly occur at the school level, occasionally occurs at the school-cohort level if there is administrative turnover.

5. Unobservably superior distribution of teachers causes a larger share of students in honors. Mostly occurs at the school level, possibly at the school-course level, rarely at the school-cohort (slight teacher turnover), more likely at the school-cohort-course level (one or two teachers turning over could induce changes in the share of students in honors).

   How do different specifications address these endogeneity concerns:

6. School fixed effects: removes all three mechanisms at the school level, but leaves them at lower levels.

7. School-cohort fixed effects: removes all three mechanisms at the school and school-cohort level, but leaves them at the school-cohort-course and school-course level.

8. School-course fixed effects: removes all three mechanisms at the school and school-course levels, but leaves them at the school-cohort and school-cohort-course levels.

9. Using past share honors as an instrument: retains school-level mechanisms, but removes school-cohort mechanism and school-cohort-course mechanism. Useful if the between school variation is cleaner. Also if there is inertia in changing policies, so that the policies predict future policies even when the conditions that spawn the policies disappear. Could use a past school-average share across courses or a past school-course share. The former removes the between course variation, while the latter preserves it.

10. Share of classrooms as an instrument for share of students: lose a lot of variation, but lumpy adjustment might cause large changes in share honors from a small change in underlying student ability distribution. This is perhaps what the school chooses, rather than the students/parents. Maybe the student/parent component is particularly likely to generate an endogeneity problem.

11. Bartik approach: Use predicted share honors as an instrument. Several versions: Use school's historical shares of students in each quintile to predict the share in honors. Or use the school-cohort's historical share of students in each quantile to predict the share in honors. This removes some of the principal discretionary variation that might be correlated with principal quality. But it might still contain some if that principal quality is causing the sorting of students of different quintiles across schools.

3 Introduction

- Decide whether restricted or unrestricted cubic is the baseline specification.
- Discuss the value of aggregating?
- Highlight the importance of large scale data: small experiments do not provide sufficient sample size to estimate nonmonotonic effects by type. Use this to motivate the source of data, and perhaps then talk about the specification.
- Need to create confidence bands for cubic estimates.
- Emphasize that small per-student gains are nonetheless consistent with large aggregate policy gains.

4 Model

- Need to suppress the stj notation in the model, focus attention on a given school-course-year.
- Should we call honors share "honors fraction" instead, since we use f for the notation?

5 Model Redraft Outline

1. Why do we include a model?
    - Justify aggregation to school-course-year-quantile level.
    - Provide guidance for the estimating equation.
    - Motivate the treatment effect functions of interest that we estimate as inputs to a planner's problem.
2. Introducing the administrator's problem

- Justify why the choice of f is within the purview of the administrator
- Introduce the maximand of interest and highlight the necessary inputs.

3. Introducing Test score production
    - Key role of additive separability
    - Justify aggregation.
4. Optimal choice by students and its sorting implications.
    - Highlight the value of aggregation and help interpret what the treatment effect parameters capture.
    - Show how endogenous sorting can be accommodated with few assumptions about its nature.
    - Highlight the importance of comparing comparable schools.

6 Research Question and Motivation

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

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

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

There are three essential challenges to estimating the impact of changing the intensive margin of honors selectivity. First, much like other school policy interventions, the expected per-student achievement impact of changing the size of the honors track is likely to be small relative to all of the other student, teacher, and school inputs that affect achievement. Thus, the amount of variation necessary to obtain sufficient power to distinguish treatment effects from alternative honors track selectivity is daunting, particularly when there are strong theoretical reasons to expect heterogeneous and non-monotonic effects from increased selectivity. In particular, the onerous sample and specification requirements generally preclude the use of small scale experiments and narrowly defined instrument variables that would otherwise provide credible identification.

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

The North Carolina administrative records we use are particularly well suited to aid us in overcoming all three challenges. The data contain histories of elementary and middle school test scores for millions of public high school students from 1995 to 2013. In addition, the data feature statewide course-specific tests in eleven high school courses, of which we focus on six for which tracks are easily inferred.[1] By facilitating comparisons across schools, across school cohorts, and across courses within a cohort, these two features ensure that an enormous amount of variation in honors track sizes and subsequent achievement can be

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

harnassed to identify heterogeneity in impacts at different margins of selectivity for different student subpopulations.

Furthermore, North Carolina's accountability system provides strong incentives to principals and teachers to adhere to the curriculum tested by the statewide exams regardless of track, which mitigates concerns about misalignment between the content taught versus tested in each track.[2] To further ensure comparability of test score performance, we drop from our sample school-course combinations featuring Advanced Placement (AP) or International Baccalaureate (IB) tracks, since schools and teachers face competing incentives to adhere to alternative curricula, and focus exclusively on courses featuring only honors and regular tracks.

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

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

We then aggregate to the school-course-year-preparedness quintile level, which sidesteps the selection problems associated with individual student's choice of track that has been the focus of much of the tracking literature. We also restrict the sample to schools with typical distributions of student past performance, so that the regular and honors peer environments associated with a given honors fraction are likely to be similar across schools. We then regress test scores on a cubic function of the fraction of students in honors in the associated school-course-year combination, along with our full set of controls. To account for heterogeneity in impact, separate cubic coefficients are estimated for each quintile of a regression index of student preparedness based on past test scores.

Thus, validity of our baseline estimates requires that, conditional on our full set of con-

[2]Among other features, state test performance contributes at least 20% of a student's GPA regardless of track in North Carolina.

trols, the variation in the share of a course's students that chooses the honors track is unrelated to other unobserved school, teacher, and student inputs that may affect test score performance.

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

We find that the students in the first (highest) predicted achievement quintile most benefit from honors programs that comprise 20-30% of the student body; they enjoy an expected increase of 0.07 test score standard deviations on average relative to a version of the course without tracking. The second quintile exhibits similar but smaller effects as the first, with an average test score gain of about 0.05 standard deviations (SDs) for the 20-30% range, but the test score gains for 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 variation in the size of the honors program, but does exhibit small gains of about 0.025 SDs relative to the absence of tracking when the share of students in honors is between 20 and 30%. The fifth quintile does not exhibit any statistically or economically significant gains from any exclusiveness 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 tracks 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. Furthermore, enough schools and courses feature suboptimal honors selectivity so that if all schools switched from their current honors program size to the optimal size, we predict that North Carolina high school students would gain an average of 0.02 test score SDs. The 20-30% range for the share of students in honors still maximizes the weighted average performance and delivers sizable gains relative to no tracking even with a weighting system that weighs the achievement gains of quintiles 1, 2, 3, and 4 at 20%, 40%, 60%, and 80% of those of quintile 5, respectively. For honors shares

greater than 30%, the benefit of having more students placed into the honors program seems to be more than offset by the cost of having both the regular and honors track decrease their average student quality and the level of instruction.

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

Our contribution to the tracking literature is to quantify the impacts of changing the intensive margin of honors track selectivity in a context where students self-select into tracks conditional on capacity constraints implicitly set by school administrators. Other papers have evaluated the extensive margin choice of whether to have any tracking, in several cases exploiting experimental or quasi-experimental variation. These papers generally do not analyze the size of the honors track when it exists. Some of these papers have found they help the top students and hurt the bottom students (**?????**). Others have found they do not hurt any students (**????**) or have small or insignificant effects (**??**). Our results suggest that these seemingly contradictory results might potentially be reconcilable if the different papers feature samples of schools with different mixes of honors enrollment shares.

? represents the rare paper in this literature that incorporates an explicit role for honors track selectivity. The authors build a structural model that includes an administrator's 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. However, while their approach permits a broader welfare analysis, it also requires strong assumptions to simultaneously identify parameters governing tracking in combination with other preference and technological parameters governing other choices in the model. Furthermore, they focus on elementary schools, and their tracking data are nowhere near as rich as the North Carolina administrative data; the authors are forced to infer the track based on variation in teachers' self reports of the quality of their students, which could simply reflect sampling error rather than tracking per se.

A second strand of the literature considers the effect on an individual student of moving into an honors or gifted track, either using regression discontinuities (**?**) or propensity score matching (**???**). These papers generally find that enrolling in advanced tracks improves test

scores for the marginal students they consider, with (**?**) Our estimates combine the effects on the marginal students with the accompanying effects of diluting the honors track and reducing the peer quality in the regular track. Our results suggest the impact of honors is not limited to just the marginal students, since students whose past test scores strongly suggest they will be inframarginal are still affected by changes in the selectivity of the honors track.

Finally, this paper also contributes to the much larger literature considering peer effects on academic achievement. While our approach does not isolate the contribution of peer effects, they are likely to be one of the driving forces for our results. **??** found that having better peers improved outcomes for students across the ability distribution. **?** found that improved peer quality increases academic performance through both congnitive and non-cognitive mechanisms, such as study time. **?** 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. Our results are consistent with theirs, since we find that top students gain most from small honors programs, where the peer quality is presumably high, and bottom students are relatively unaffected by small honors programs, suggesting that they are relatively insensitive to peer effects from top students. By employing adding additional assumptions about student assignment, **?** was able to separately identify peer effects, and similarly finds that changing the fraction of students in honors induces different peer effects which differ by the type of student affected.

The remainder of the paper will be structured as follows: Section 7 presents a theoretical model that governs the administrator's implicit choice of the size and/or selectivity of the honors track. Section 8 then describes the data, Section **??** lays out the empirical approach, Section 10 reviews the results, Section 11 provides several robustness checks, and Section 12 interprets the findings and concludes.

7 Model

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

7.1 The Administrator's Problem

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

Since the administrator can adjust these incentives separately for each course and cohort, we consider the administrator's problem for an unspecified course and year, and suppress any dependence of the inputs on course and year. Let f denote the chosen fraction of students in honors. Let θ_q denote the preference weight that the administrator gives to the performance of subgroup q, and let W_q denote the share of students in subgroup q among the chosen course and cohort. While these subgroups could be arbitrary combinations of predetermined observable student characteristics, in our empirical work we will use quintiles of predicted student performance based on students' test score histories. 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. Finally, let $E[Y_i(f)]$ and $E[\overline{Y}_q(f)]$ capture the expected test score of student i and the expected mean test score of students in subgroup q, respectively, as a function of the chosen honors fraction f. Assuming that administrators seek to maximize some weighted average of student performance, we can

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

write the administrator's problem as:

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

This formulation makes clear that the principal does not need to predict exactly which students will switch track when the chosen honors fraction changes nor the impact from any given individual from switching track or experiencing a more selective track. Rather, the principal only needs to understand how shifting f changes the mean performance of each subgroup after classroom sorting re-equilibrates. This insight motivates our approach of aggregating over individual track choice and comparing mean outcomes of different subpopulations under different tracking regimes. We make this point more explicit in the next subsection.

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

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

7.2 Test Score Production

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

$$Y_{istj} = d^h_{istj}\tilde{h}(q_{istj}, \epsilon_{istj}|\vec{q}_h, \vec{\epsilon}_h) + [1 - d^h_{istj}]\tilde{r}(q_{istj}, \epsilon_{istj}|\vec{q}_r, \vec{\epsilon}_r) + X^O_{istj}\beta^O + X^U_{istj}\beta^U + \mu_{istj}. \tag{3}$$

The students choice of track is represented by the indicator variable d^h_{istj}, with a value of 1 signifying enrollment in honors and a value of 0 signifying enrollment in the regular track. The functions $\tilde{h}(q_{istj}, \epsilon_{istj}|\vec{q}_r, \vec{\epsilon}_r)$ and $\tilde{r}(q_{istj}, \epsilon_{istj}|\vec{q}_r, \vec{\epsilon}_r)$ capture shifts in achievement from taking the honors and regular tracks, respectively. These shifts are functions of the student's own inputs, which are partly predictable based on the student's observable subgroup q_{istj} but also depend on an unobservable idiosyncratic component ϵ_{istj}. ϵ_{istj} captures deviations in expected performance, perhaps due to accumulated skills or effort unaccounted for by subgroup. Such deviations vary not just across students but within students across school-course-year combinations. Importantly, the impact of the track choice on achievement also depends on the peer environment within that track, which is reflected in the dependence of

the $h(*)$ and $r(*)$ on the vectors $(\vec{q}_h, \vec{\epsilon}_h)$ and $(\vec{q}_r, \vec{\epsilon}_r)$ capturing the subgroups and idiosyncratic contributions of other members of the honors and regular tracks. This flexible formulation of track effects acknowledges that students' production in the classroom will be affected by how the material matches with their ability and how the peer environment interacts with their own ability and effort. Track-specific teacher inputs are implicitly assumed to be functions of the kinds of students selecting into the track in a given school-course-year.

X^O_{istj} and X^U_{istj} capture other observed and unobserved student, school, or course inputs, respectively, that affect i's learning, while μ_{istj} captures measurement error that causes the test score to fail to reflect the student's learning in the chosen course. Importantly, by imposing that these inputs are additively separable from the inputs that enter the track-specific functions $h(*)$ and $r(*)$, we have assumed they have the same impact on test scores regardless of track. This implicitly requires that the standardized tests used to assess knowledge in each course do not depend on the track chosen, which is true for the honors and regular tracks we consider in our data.[4] While somewhat restrictive, the additive separability assumption implies that these inputs are irrelevant to the administrator's tracking problem; we can rewrite achievement in terms of the difference between performance in the chosen track and performance in a pooled version of the course with no tracks :

$$\Delta Y_{istj} = d^h_{istj} h(q_{istj}, \epsilon_{istj} | \vec{q}_h, \vec{\epsilon}_h) + [1 - d^h_{istj}] r(q_{istj}, \epsilon_{istj} | \vec{q}_r, \vec{\epsilon}_r) + \mu_{istj}. \tag{4}$$

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

7.3 A Simple Model of Student Track Choice

Now consider the student's choice of honors vs. regular track. Suppose that each student chooses the track that maximizes his or her test score net of track-specific effort costs, scheduling opportunity costs, and GPA boosts. Let c_{istj} capture the net difference in student i's idiosyncratic composite cost of joining the honors track h relative to the regular track r at school s in course j at time t. Next, let α_{stj} capture a component of the net composite cost difference that is common to all students in (s, t, j). Importantly, suppose that the

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

administrator has the ability to shift α_{stj} by any arbitrary amount by adjusting the relative GPA boost or homework load in the honors track.

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

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

Next, suppose that the joint distribution of the students' observed ability groups, unobserved ability components, and idiosyncratic effort/scheduling costs is given by $g_{stj}(q, \epsilon, c)$. Then we can define $\alpha^*_{stj}(f)$ as the threshold common cost component α^*_{stj} that causes a fraction f of students in the chosen school-year-course to choose the honors track. Specifically, $\alpha^*_{stj}(f)$ is implicitly defined as the solution to the following equation:[5]

$$\iiint d^h_{istj}(\alpha_{stj}, q, \epsilon, c) g_{stj}(q, \epsilon, c) dq d\epsilon dc = f. \tag{5}$$

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

Assumption 1. $g\_stj(q, \epsilon, c) \approx g(q, \epsilon, c) \; \forall \; (s, t, j) \in \mathcal{S}$

Under Assumption 1, as courses become large the threshold cost function $\alpha^*_{stj}(f)$ becomes common among sufficiently similar schools and course-year combinations within schools: $\alpha_{stj}(f) \approx \alpha_f$ for all $(s, t, j) \in \mathcal{S}$.

Furthermore, because the conditional distribution $g(q, \epsilon, c|d^h)$ also becomes common, the vectors of track-specific peers $(\vec{q}_r, \vec{\epsilon}_r)$ and $(\vec{q}_h, \vec{\epsilon}_h)$ also depend only on f (through $\alpha^*(f)$) rather than separately on s, t, or j. This in turn implies that $h(q_{istj}, \epsilon_{istj}|\vec{q}_h, \vec{\epsilon}_h) \approx h(q_{istj}, \epsilon_{istj}|f)$ and $r(q_{istj}, \epsilon_{istj}|\vec{q}_r, \vec{\epsilon}_r) \approx r(q_{istj}, \epsilon_{istj}|f)$. It also implies that the subgroup-specific probability of choosing honors depends only on f:

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

Thus, the implicit choice of f by the administrator (through $\alpha^*(f)$ can serve as a sufficient statistic for the peer composition of both the honors and regular tracks in all school-year-course combinations where this common joint distribution of ability in costs represents a

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

sufficiently close approximation. In our empirical work, we attempt to make this approximation plausible by removing schools from our sample whose students exhibit a distribution of past performance on state exams that is too far from the state norm.

However, even if the joint distribution $g(q, \epsilon, c)$ is roughly common among schools, it may not be known by any school administrator, since both ϵ and c are unobserved for each student. Thus, any given principal will have a difficult time inferring both $g(q, \epsilon, c)$ and the track-specific achievement functions $h(q, \epsilon|f)$ and $f(q, \epsilon|f)$ from data on student performance.

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

$$E[\Delta\overline{Y}_q(f)] = P(d^h = 1|q_{istj}, f)E_{\epsilon\in h}[h(q_{istj}, \epsilon|f)] + [1 - P(d^h = 1|q_{istj}, f)]E_{\epsilon\in r}[r(q_{istj}, \epsilon|f)] \tag{7}$$

where $E_{\epsilon\in h}[h(q_{istj}, \epsilon|f)]$ and $E_{\epsilon\in r}[r(q_{istj}, \epsilon|f)]$ are defined by:

$$E_{\epsilon\in h}[h(q_{istj}, \epsilon|f)] = \frac{\iint d^h(\alpha^*(f), \epsilon, c, q)h(q_{istj}, \epsilon|f)g(q_{istj}, \epsilon, c)d\epsilon dc}{\iint d^h(\alpha^*(f), \epsilon, c, q)g(q_{istj}, \epsilon, c)d\epsilon dc} \text{ and} \tag{8}$$

$$E_{\epsilon\in r}[r(q_{istj}, \epsilon|f)] = \frac{\iint (1 - d^h(\alpha^*(f), \epsilon, c, q))r(q_{istj}, \epsilon|f)g(q_{istj}, \epsilon, c)d\epsilon dc}{\iint (1 - d^h(\alpha^*(f), \epsilon, c, q))g(q_{istj}, \epsilon, c)d\epsilon dc} \tag{9}$$

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

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

8 Data & Background

We use administrative data provided by the North Carolina Department of Public Instruction for all public schools between 1995 and 2013. These data and their surrounding institutional context have several important features that make it well suited for our analysis.

First, the track associated with each high school classroom is reported for each course, both by school administrators at the beginning of the year and directly by students during assessments at the end of the year. Such dual reporting provides confidence that track is being measured correctly.[6]

Second, North Carolina required statewide standardized end-of course exams as part of 11 distinct high school courses during our sample period. Importantly, because the same exams were administered to all schools and all tracks within a school, these test scores represent a common metric by which schools choosing different selectivity of honors programs can be compared. Student performance on these exams contributes to a state-mandated minimum 20% of the students course grade, so students an incentive to perform well and teachers have an incentive to adhere to the curriculum regardless of track. Hence, in this North Carolina context, the honors track is likely to primarily represent greater depth and difficulty of covered material rather than greater breadth.

We exclude five of these courses from our sample due to either a small set of test years (Civics and Econ, Law & Politics), inconsistency in grade level (Algebra 1 is often offered in middle school rather than high school), or preponderance of Advancement Placement classrooms, discussed further below (US History and Physics). Thus, our sample consists of standardized scores from the following six courses: Algebra 2, Biology, Chemistry, English 1, Geometry, Physics, Physical Science, and US History. Appendix Figure **??**, which displays the statewide distribution of student scores of the courses in my final specification for the year 2006, reveals no evidence of any floor or ceiling effects.[7]

Table **??** examines the tracking options available in each course for school-year-courses with at least 30 student observations. For most school-year-courses, there exists an honors program, but remedial programs are rare. Furthermore, the remedial track generally accounts for a very small portion of the student body in it when it exists (see Figure **??**). Given insufficient power to detect the impact of alternative remedial track sizes, we drop remedial classrooms from our sample. We also drop classrooms from the advanced placement

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

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

and international baccalaureate tracks, because teachers in these courses may adapt their curriculum to align more with the AP exam than the NC end-of-course exam, making the latter test scores less accurate measures of learning. Thus, we focus attention on regular and honors tracks, and use the share of students attending regular and honors track classrooms in a given school-course-year who are in fact in the honors track as our main independent variable of interest, in alignment with the honors fraction f in Section 7 above.

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

Finally, the North Carolina administrative data offers a wide array of observed control variables at the school, teacher, classroom, and student levels. As emphasized in the following section, such rich controls are critical for addressing omitted variable bias stemming from correlations between the honors share and other school, teacher, and student inputs that contribute to test scores. Of particular note are histories of students' standardized test scores during grades 6-8 in math, English and (for some cohorts) science. These histories provide powerful controls capturing differential student preparedness across schools and cohorts that might both influence principal's decisions about honors selectivity and predict future student performance.

8.1 Assignment to Preparedness Quintiles and Restricting the Sample of Schools

These test score histories also provide a basis for assigning students to the observed quality types that are necessary for providing a holistic assessment of alternative choices of honors selectivity. Specifically, we assign each student to a predicted quintile in the statewide performance distribution (with quintile 1 denoting the highest predicted performance) based on the distribution of students' regression indices from a regression of test scores in the sampled high school subjects on grade 7 and 8 English and math scores. Importantly, we allow the coefficients on these past scores to be course-specific, so that the same student may be assigned to different quintiles for different courses if their past performance indicates

different relative strengths in the skills required by these courses.[8] For the sake of brevity, henceforth we refer to these statewide predicted quintiles merely as quintiles for the rest of the paper, and will be explicit in the few cases in which within-school student rankings are used as the basis for assignment to a quintile.

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

However, a less selective honors track may result in a considerably different ability distribution among students choosing the track at an extremely privileged school relative to a school with few resources and struggling families. To gauge the scale of the problem, Appendix Figure **??** looks at how many quintiles students would need to shift, on average, in order for a school to match the statewide (uniform) distribution of predicted quintiles. While the majority of schools appear to have nearly uniform distributions, there is a substantial right tail of schools with substantially skewed distribution. Thus, we restrict the sample to schools which require fewer than 0.5 quintile changes per student to match the uniform distribution, which removes about 30% of the observations from the original sample. Appendix Figure **??** shows the histograms of the six schools with the aforementioned metric closest to and less than one half. While this sample restriction ensures plausible comparability among schools, it may limit external validity of our estimates to schools with very low or very high student past performance. As a robustness check, we also consider a specification where the above metric for the spread of student quality is less than one third. Appendix Figure **??** displays histograms of the six school-courses where the average number of quintile shifts required for a uniform distribution are closest to a third.[9]

[8]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.

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

8.2 Summary Statistics

If Assumption 1 holds, each principal is perfectly informed about $E[\overline{Y}_q(f)]$, and each has the same preference weights θ_q, then each principal's optimal choice of f to solve (1) would be the same, and there would be no identifying variation in f. Appendix Figure **??** allays this fear by displaying the distribution of honors shares for the six courses in our final sample among school-year-courses with honors programs. Every subinterval between 0.1 and 0.6 shows frequent use in all six courses, and Chemistry features a nontrivial share of school-year combinations with more than 60% of students in honors. This suggests that administrators either vary in preference weights θ_q or in differential misperceptions about $E[\overline{Y}_q(f)]$. Given the dearth of convincing evidence from the literature on the particular tradeoffs associated with different honors track sizes, such misperceptions are not surprising.

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

Table **??** provides descriptive statistics at the school-course-year for the remaining sample of high schools by selectivity of the honors track. Relative to school-course-years without tracking, those with smaller honors tracks ($< 35\%$ of students) have very similar distributions of student achievement, teacher credentials, and student demographics and parents' education. The one major difference is that schools with smaller enrollment are more likely not to offer an honors track. Relative to both school-course-years without tracking and with smaller honors tracks, those with larger honors tracks ($> 35\%$ of students) tend to have students with slightly more educated parents. Thus, at first blush it seems that much of the variation in honors fractions is not directly tied to the composition of students or teachers at the school.

Figure **??** plots the average honors enrollment rate for bins of the coursewide honors fraction separately by within-school (rather than statewide) quintile of predicted performance alongside the enrollment rate one would expect if students were perfectly sorted to tracks based on their relative predicted performance. Perfect sorting on predicted performance would result in line with a slope of 5 within the interval of honors fraction corresponding to the chosen quintile ([0,.2] for quintile 1, [.2,.4] for quintile 2, and so on) and a flat line with

zero slope elsewhere. The final cell in Figure **??** shows the pooled distribution of honors share among all school-course-years in the sample.

On the one hand, students in top quintiles unsurprisingly enroll in honors at much higher rates than students in other quintiles. Nonetheless, the plots of observed honors enrollment patterns reveal quite imperfect sorting, suggesting that unobserved ability and heterogeneous effort costs do play an important role in track choice.[10] For example, a course with 20% of students in honors tends to be chosen by only 60% of top quintile students (rather than the predicted 100%) and by 25%, 15%, 5%, and 1% of students in quintiles, 2-5, respectively, (CHECK NUMBERS!). Similarly, a course with 60% of students in honors still has 20% of students in quintile 5 enrolling in honors while 20% of quintile 2 students enroll in the regular track. For quintiles 2 and 3 in particular, these unobserved sorting factors play a large role in enrollment, as both quintiles have significant enrollment rates for all shares of students in honors.

Figure **??** plots the average contemporaneous test score performance in statewide test score standard deviations for bins of the share of students in honors, separately by quintile. Interestingly, the average performance within each preparedness quintile is at or near its peak when the share of students in honors is around 40%.[11] However, in order to verify that this finding reflects a true pareto-optimal honors share rather than a spurious correlation between honors fraction and other school and student inputs, we now describe our more rigorous estimation procedure.

9 Empirical Approach

9.1 Baseline Specification

Our primary specification is an aggregated version of the education production function (3) from Section 7.2. Recall from Section 7.3 that the objects of interest, quintile-specific treatment effect functions of the honors fraction ($E[\Delta\overline{Y}_q(f_{stj})]$), are aggregate objects that only vary at the school-year-course-quintile level. Furthermore, since the control variables X^O_{istj} enter linearly in f_{stj} aggregating to the school-year-course-quintile level Thus, because the control variables X^O_{istj} enter linearly and are assumed to be additively separable from $E[\Delta\overline{Y}_q(f_{stj})]$, we can estimate the parameters of interest in (3) at the school-year-course-quintile level without introducing any bias and with minimal lost efficiency. Indeed, such aggregation allows us to avoid selection problems from individual track choice. Thus, our

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

[11] Note that 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.

primary specifications all take the following form:

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

Each specification implements $E[\Delta\overline{Y}_q(f_{stj})]$ as a set of quantile-specific, flexibly parameterized functions of f_{stj}, the fraction taking the honors track among all students taking course j in school s in year t, with the chosen functional forms varying across specifications. Our baseline specification assumes that each quantile's $E[\Delta\overline{Y}_q(f_{stj})]$ takes the form of a restricted cubic function:

$$E[\Delta\overline{Y}_q(f_{stj})] = \sum_{q'} 1(q = q')[\gamma_{q'}^{lin} f_{stj} + \gamma_{q'}^{sq} f_{stj}^2 - (\gamma_{q'}^{lin} + \gamma_{q'}^{sq}) f_{stj}^3] \tag{11}$$

The coefficients in equation 12 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.[12] This functional form permits a wide variety of shapes while still exploiting the efficiency gains from summarizing a function with two parameters. We also present results from an unrestricted cubic specification as a robustness check in Section 11. Importantly, the coefficients $\vec{\gamma}^{lin} = \{\gamma_1^{lin}, ..., \gamma_5^{lin}\}$ and $\vec{\gamma}^{sq} = \{\gamma_1^{sq}, ..., \gamma_5^{sq}\}$ are quintile-specific in order to capture heterogeneous effects by levels of student preparedness.

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

Our baseline specification pools all the variation in the honors fraction f_{stj} that occurs between schools, between years within schools, and between courses within school-year combinations to generate maximally precise estimates of the parameters $\vec{\gamma}^{lin}$ and $\vec{\gamma}^{sq}$. On the one hand, there are plausible sources of exogenous variation at each of these levels. For example, smaller schools may not be able to support the multiple number of classrooms per course that tracking requires, and the school size thresholds beyond which additional classrooms can be supported may not otherwise affect student outcomes (beyond simple class size effects that can be easily controlled for). Similarly, due to differential parental pressure,

[12] 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.

personal pedagogical beliefs, or accountability pressure, principals may differentially weigh performance by different quintiles or have incorrect beliefs about the impact of tracking for reasons unrelated to any of the other unobserved inputs affecting these students performance, leading to different chosen honors fractions. Switching costs from new course preparation for certain teachers may cause schools not to track even when other similar schools do so, perhaps because of differences in the past course histories of their teachers.

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

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

Unfortunately, observable variables that isolate only the exogenous sources of variation are either not available or generate instruments that are too weak to detect the heterogeneity in achievement impacts across the student ability distribution. Since we estimate the model via ordinary least squares, in order for our baseline estimates to be unbiased, unobserved inputs contained in the error term must be uncorrelated with the honors share f_{stj} as well as its square and cube, conditional on the controls X_{stjq} and Γ_{stjq}.[13] Thus, our baseline specification relies heavily on the richness of the North Carolina administrative data to provide a set of powerful controls that absorbs the most plausible sources of endogeneity.

To address endogeneity from student composition, in our baseline specification the vector X_{stjq} contains student ability and preparedness measures (mean test scores in grade 7-8

[13]Specifically, we assume $E[\mu_{stjq} f_{stj} | X_{stjq}\beta^X, \Gamma_{stjq}\beta^\Gamma] = 0$, $E[\mu_{stjq} f_{stj}^2 | X_{stjq}\beta^X, \Gamma_{stjq}\beta^\Gamma] = 0$, and $E[\mu_{stjq} f_{stj}^3 | X_{stjq}\beta^X, \Gamma_{stjq}\beta^\Gamma] = 0$.

math and english interacted with quintile), share with gifted status), student demographics (race shares, ESL shares), and family socioeconomic indicators (parental education categories, share with free/reduced price lunch status). Note that introducing a variety of such measures as aggregate shares at school and school-course-year-quantile levels rather than individual-level controls also implicitly controls for unobserved school and cohort characteristics by potentially spanning the common amenity space that lures certain observably and unobservably kinds of students to the school or course within the school ((**?**)).

To address endogeneity from teacher inputs and remaining school inputs beyond those that affect student composition, X_{stjq} also includes proxies for teacher quality (experience, certification scores, degree/license status), and controls for school size, and statewide accountability status. We also control for class size, an important input that is likely to be affected by otherwise idiosyncratic changes in honors shares. The full list of baseline controls and their sample means are provided in Appendix Table XX. All controls are interacted with the full set of course indicator variables to allow differential predictive power in different courses. Finally, we include in Γ_{stjq} a full set of year-course-quintile fixed effects, which removes potential bias from changes in statewide course curricula or the relative difficulty of standardized test questions that target different parts of the ability distribution that may be correlated with statewide trends in honors fractions.

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

The first alternative specification adds a set of school fixed effects to Γ_{stjq}, so that the parameters of interest are only identified by differential changes in achievement across cohorts and courses within schools. While many of the most pressing endogeneity problems involve greater honors selectivity causing or responding to student sorting among schools, introducing school fixed effects also generates noisier estimates, since between school variation accounted for 39% of residual identifying variation net of controls in our baseline specification.

The second alternative specification we consider uses the honors share of the previous cohort in the same school-course combination, along with its square and cube, as instruments for the corresponding contemporaneous share and its square and cube. The exclusion restriction for this IV specification requires that the past share of students in honors affects current test scores only through inertia in the share of honors over time conditional on controls. This IV approach purges estimates of any endogenous honors share response to unobservable changes in cohort quality within a school. Implicitly, this specification puts greater emphasis on between-school and within-school/across-course variation at the expense

of time-series variation.

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

While none of these alternative specifications is intended to allay all fears about bias in isolation, collectively they can potentially provide considerable reassurance if results are consistent across all of these specifications. After all, if substantial endogeneity biases exist, they would need to operate with the same force (relative to the exogenous variation) at each level of variation *and for each quintile of student preparedness* in order to generate such consistency. Put another way, our flexibility in allowing separate cubic functions of honors fraction for each quintile also provides more opportunities for sizeable endogeneity biases to reveal themselves through distinct results patterns across specifications that magnify or reduce the role these sources of endogeneity play in driving results.

We cluster standard errors at the school level in each specification, both to be conservative and because we expect considerable autocorrelation in errors across course-years from the same school. In addition, each specification weights observations 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 within-school-year shares would prioritize the efficacy of administrators' actions at large schools over smaller schools. Given that we are interested in providing inputs to principals of all school types, we prefer to weighting schools rather than students equally. This weighting scheme also acts as additional correction beyond clustering for school-size-related heteroskedasticity in the error. As per the recommendations of (**?**), we have also re-run each of these specifications with weights proportional to the number of students in the school-year-course quintile combination so as to weight each student equally. Point estimates and standard errors are similar for the different weighting schemes[14]

[14]These specifications are available upon request from the authors.

10 Results

10.1 Quintile Treatment Effects

The red lines of Figure XX display predicted values of treatment effects on acheivement for a dense grid of potential honors fractions from our baseline restricted-cubic specification that pools all sources of residual variation in the honors fractions among school-year-course combinations. Note that all predicted values capture treatment effects for alternative honors fractions relative to an absence of tracking, which has been normalized to zero. Dashed blue lines indicate the upper and lower bounds on 95% confidence intervals (CI) that were created by using the delta method to convert the variance-covariance matrix associated with point estimates for the cubic parameters $\vec{\gamma}$ into confidence intervals for each predicted value along the grid.[15] The bottom right cell in the figure displays the support of the honors share distribution for school-year-courses that feature an honors track. Since there is limited support for honors programs with shares greater than 65% and between 0 and 15%, predicted values in these ranges are primarily driven by functional form assumptions and should be viewed skeptically.[16] Column XX of Table XX provides the predicted values and their standard errors for the baseline specifications for each quantile for several candidate honors fractions that underlie Figure XX, while Appendix Table **??** provides the underlying parameter estimates $\vec{\gamma_q}$ for each quantile q from different specifications in this section as well as in Section 11.

Starting with quintile 1, we observe that 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 relative to the absence of tracking. This is similar to the predicted increase in student achievement associated with switching from the median teacher to a 76^{th} percentile teacher ((**?**)). However, these gains quickly disappear as the honors fraction increases beyond 30%. Since the vast majority of quintile 1 students will enroll in honors if it contains at least 30% of their cohort, the sharp decrease in gains as honors becomes more selective is likely due to the dilution in peer quality within the honors track. **?** found that high achieving students are especially sensitive to peer effects, potentially justifying why quintile 1 experiences such a sharp decrease as the share of students in honors is increased.

Students in quintiles 2 and 3 also particularly benefit from smaller honors programs.

[15] Note that such pointwise confidence intervals differ from 95% joint confidence bands, which delineate an area that spans the entire treatment effect function with 95% probability. Such confidence bands would be necessary for evaluating joint hypotheses involving predicted values over continuous range of honors fractions, such as whether there is any nonzero honors fraction that makes quantile 2 worse off than the absence of tracking.

[16] Note that the gradient in the delta method is equal to zero when the share of students is one or zero.

Relative to the absence of tracks, the gains from the existence of an honors track rise until a peak gain of about 0.05 SDs and 0.04 SDs, respectively when 40% of all students are choosing honors. Interestingly, this peak occurs at a fraction where large shares of students in these quintiles are near the margin of choosing honors: around 55% of quintile 2 students and 35% for quintile 3 students generally enroll in an honors track that serves 40% of the cohort, with these shares continuing to rise significantly when the share of students in honors increases beyond 40%.

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

Decomposing these competing mechanisms to isolate how each incremental expansion of the honors track affects marginal students, inframarginal honors track students, and inframarginal regular track students within each quintile would require strong assumptions on the degree to which unobservable ability vs. scheduling costs is driving students' selection of track. Indeed, the appeal of our approach is that it can provide the policy-relevant inputs for administrators without requiring questionable assumptions about student sorting to tracks. Thus, we leave such a decomposition to future research.

Quintile 4 students seem to be quite insensitive to the size of the honors track. Equivalence of a two track menu with a trackless course can only rejected with 95% confidence for a narrow range in which the share of students in honors less than 30%. The point estimate at the peak near 25% is about .03 SDs.

Quintile 5 exhibits only small, statistically insignificant gains from small honors programs, and begins to experience losses relative to a no tracking regine once the honors program grows beyond 40%. These results are consistent with the peer effect literature that has found that lower achieving students are the least sensitive to the positive peer effects from the highest ability students (**????**). Although having a small honors program decreases the average peer quality for the overwhelming majority of bottom quintile students who don't enroll in honors (over 90% remain in the regular track with a 40% cohort-wide honors share), the compositional changes may be offset by a better paced class. However, perhaps when the honors program grows beyond 40%, the bottom quintile students who do not enroll in honors (still around 80% when the cohort-wide percent in honors is 60%) no longer share the

classroom with the middle tier students with whom they might otherwise profitably interact.

To assess the sensitivity of the results to the source of variation, we next consider the three alternate specifications introduced in Section 9 whose estimated treatment effect functions are presented in Figures **??**, **??**, and XX. Recall that the alternate specifications are identified by different subsets of the variation identifying our baseline model. Figure **??** corresponds to the school fixed specification that isolates variation in honors policies that either change over time or vary by department within a school. The school fixed effect specification yields quite similar results to the baseline specification across all quintiles, both in shapes and magnitudes.

Figure **??** presents results from the first IV specification, which uses lagged course-specific honors shares as instruments for current honors shares. It seeks to remove cohort-specific variation at each school while leaving both stable between-school and stable between-course within-school differences in honors selectivity. It is motivated by the idea that school and department administrators may have idiosyncratic preferences or beliefs about honors efficacy that systematically shape their default choices of honors selectivity across years. This specification yields point estimates that are noisier but also slightly larger in magnitude than the baseline specifications.[17] While the larger magnitudes could simply be sampling error or a greater upward bias, another possible explanation is that honors shares may be reported with error that is corrected by the IV specification, suggesting that the estimates from the baseline specification may be attenuated.

The third alternative specification alters the IV approach by using a common (mean) lagged honors fraction as the instrument for contemporaneous honors fractions across all courses in the same school-year. This specification removes systematic between-course variation in honors shares as well as transitory cohort-specific variation, leaving only persistent differences in schools' tendencies to have larger or smaller honors tracks. Figure XX shows that this second IV approach yields results that are similar to (albeit noisier than) the original IV approach, demonstrating that general patterns do not hinge exclusively on the exogeneity of the residual between-course variation. Perhaps most importantly, though, the baseline specification and all three alternative specifications yield similar qualitative patterns: 1) students in the top quintiles benefit significantly from honors programs with fewer than 30% of the student body in them; 2) students in the 2nd and 3rd quintiles benefit form honors programs with 20-40% of the student body in them; 3) Students in the 4th quintile are relatively unaffected by changing the fraction of students in honors, with potentially

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

small gains from small honors programs; and 4) students in quintile 5 are on average unaffected by honors programs with less than 40% of the student body in them and hurt by honors programs with more than 40% of the student body in them. As emphasized above, such consistency is unlikely to occur if endogeneity were driving the results, since different sources of endogeneity would need to cause the same pattern of bias across the interval of honors shares for all five quintiles of the preparedness distribution.

Interestingly, our results show that honors tracking programs are not a zero sum game. Small honors programs (between 25% and 40%) provide a Pareto improvement across quintiles relative to large honors programs (> 40%), with some quintiles exhibiting large gains. One can potentially reconcile our results with papers finding that introducing tracking does not harm any students if those students are primarily at schools that have small honors programs ((**???**)). Similarly, one can also potentially reconcile our results with papers finding that honors programs help top students and hurt bottom students if those papers sampled a greater share of schools with larger honors programs ((**????**)).

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.

10.2 Administrator's Problem

Given estimates just presented of the quantile specific treatment effect functions $\{\hat{E}[\Delta\overline{Y}_q(f)]\,\forall\, q \in [1,5]\}$ functions, we are prepared to reconsider the administrator's problem (2) from Section 7.1. Recall that solving for the optimal choice of honors selectivity also requires supplying weights $\{\theta_q\}$ capturing the relative importance the administrator places on achievement gains from each quintile of the student preparedness distribution. In particular, we consider two sets of weights. The first set weighs all quintiles equally ($\theta_q = \frac{1}{5}\ \forall q$), while the second set strongly prioritizes bottom quintiles so that quintiles 1, 2, 3, and 4 are weighted at 20%, 40%, 60%, and 80% of quintile 5 respectively ($\theta_q = \frac{q}{15}\ \forall q$).[18]

The left panel for Figure **??** shows the average net student gains as a function of the honors fraction under equal weighting of quintiles, based on the estimates from the baseline specification.[19] The maximized gain of 0.04 SDs relative to the absence of honors programs occur when honors tracks contain between 20 and 30% of students. The right panel of Figure **??** displays weighted average gains with the second set of weights that prioritize students in bottom quintiles. Notably, the maximum weighted average gain still occurs at honors programs with enrollment shares between 20 and 30%, with a weighted average impact of

[18] More weighting schemes are available upon request from the author.
[19] Confidence intervals for this section are also created using the Delta Method.

0.03 SDs. More generally, smaller honors programs dominate larger ones for any weighting scheme that places at least 10% of the weight on ech of the 5 quintiles. Further increases in the share of students in honors beyond 35% generate consistent decreases aggregate achievement gains for every weighting scheme over the remaining support of the data. The remarkable robustness of the optimal honors program size across weighting schemes is driven by gains for the top 60% of students from small honors programs and the lack of effect small honors programs have on students in the bottom 40% of preparedness.

The optimal size for an honors track is also robust across specifications. Figure **??** displays the average effect for the three alternate specifications under both weighting schemes. The school fixed effect specification, on the left side of Figure **??**, has a smaller maximized weighted average gain, but the optimal share in honors remains between 20 and 30%. The course-specific IV specification presented in Figure **??** has larger point estimates for the weighted average gain, but the same optimal share of honors. The pooled IV specification has XXXXX.

A 0.04SD aggregate gain from introducing an honors track serving 25% may seem relatively small; it would move a student at the statewide median to the 51.6th percentile. However, it would apply to all students in the cohort for every course in which tracking is introduced. Also, the small value may be misleading given that the lion's share of achievement variance that is determined by parents, teachers, and previous schooling, and is thus beyond the control of the high school.

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

Under this assumption, students at schools that introduced an optimally sized honors tracks could expect their earnings at age 28 increase by an average of 0.4% compared to if their school had no honors tracks for each core course in my sample.[20] For a high school

[20]This calculation assumes for simplicity that all students would have the 2018 median income of $36910 at the age of 28 in the absence of tracking, and that test score gains from each subject can be translated to earnings gains and then aggregated across subjects.

class of 100 students near the age 28 median income, this implies an increase in aggregate age 28 earnings of over $88,000. If other courses not tested, such as English classes other than English 1, had similar effects then this estimate could be much larger.

Of course, many schools already feature tracks near the optimal size for most of their courses. However, there remain a substantial share of school-year-courses in our sample that either do not use tracking or feature honors track sizes well outside the optimal range. If all schools in our sample switched from their current honors program size to an honors program with 20 to 30% of the student body in it, our estimates suggest that the average North Carolina student would create a test score gain of over 0.02 SDs (about the same amount as switching from the median teacher to a 57$^{\text{th}}$ percentile teacher (**?**)). Since North Carolina averages about 100,000 students per grade per year, this corresponds to an aggregate increase in earnings for 28 year olds state wide of over $44 million.

Clearly, such back-of-the-envelope calculations are wildly speculative; for example, they ignore general equilibrium effects in the labor market as well as the substantial costs with staffing multiple tracks at small schools.[21] Nonetheless, they serve to highlight the possibility that small student gains from a superior tracking system can aggregate to very large earnings contributions when combining effects across all courses, schools, states, and years.

11 Robustness Checks

In order to maximize power, all of the results presented to this point have imposed that each quantile's expected achievement follows a restricted cubic function of the fraction of students in the honors track that takes on zero values at both ends of the unit interval. However, to demonstrate that our main findings are not driven primarily by assumptions about functional form, here we present results from three alternative specifications for the shape of $E[\Delta\overline{Y}_q(f_{stj})]$. Figure **??** plots a flexible semi-parametric specification that replaces the cubic specification with a full set of interactions between student preparedness quintiles and quintiles of the fraction of students in honors:

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

Despite the greater imprecision, one can clearly see the same qualitative patterns for each quintile as Figure **??**. Specifically, for quintiles 1-4, expected gains compared to no tracking are generally above 0 for honors shares between 0 and 20%, then rise further between 20-40%

[21] Note that a full welfare analysis also requires incorporating the effort costs paid by students. Thus, such an analysis would require an array of additional assumptions that we are reluctant to impose. See **?** for an example of a complete welfare assessment.

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

Next, we consider relaxing the restriction that 100% of students in honors is equivalent to 0%. The motivation for relaxing this assumption is the possibility that a designation of "honors" connotes higher standards and a slightly more rigorous curriculum even when all the students are the same. Appendix Figure **??** displays the results from an unrestricted cubic specification that fits three parameters per quintile. The shapes of the conditional expectation functions are quite similar over the range between 0 and 70% honors that spans nearly the entire support of the data, so that the specifications only differ in their extrapolations to never-observed honors shares above 70%.

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

$$h^q(z_{stjq}) = \gamma_q^{lin} z_{stj} + \gamma_q^{sq} z_{stj}^2 - (\gamma_q^{lin} + \gamma_q^{sq}) z_{stj}^3 + \gamma_q^{indicator} \mathbb{1}_{(z_{stj} \in (0,1))} \tag{13}$$

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

Finally, we also consider two additional specifications that trade reduced precision in exchange for arguably better isolating exogenous variation. First, we employ a specification that uses the share of classrooms that are assigned to the honors track as an instrument for the share of all students who take the honors track, following the "Maimonides rule" identification strategy of (**?**) and others. Essentially, the high per-pupil staffing cost of offering class times with very few students may limit the set of viable honors fractions a school can choose.[22] Such granularity may cause relatively small differences in cohort sizes to experience substantial arguably exogenous differences in honors shares (conditional on class size controls). Appendix Figure **??** displays the treatment effect function from this additional IV specification. Again, the basic patterns remain the same for all quintiles.

Second, we augment our baseline specification with a full set of school-year combination

[22]For example, a school with around 75 students in a cohort may have too many students for two classes and two few for four classes, so that the only feasible shares of honors classes are 0, .33, and .66. A larger cohort of 90 students might force the school to allocate four classes, leading to honors class shares of 0, .25, .5, or .75.

fixed effects, so that estimates are identified exclusively by comparisons in relative performance across courses featuring different honors share within cohorts. These fixed effects are likely to remove almost all bias caused by student sorting, since these courses are being populated by nearly the same set of students.[23] Thus, any remaining bias would require either that the relative honors share responds to particular cohorts' unobserved mean comparative advantage in some subject (likely to be negligible) or that it responds to differential unobserved mean teacher quality or track-specific experience across courses (

Finally, it is possible that the school-courses chosen in section 4 are not sufficiently similar in their joint distributions of student abilities and costs to satisfy Assumption 1 and thus make the peer environment comparable in different schools featuring the same honors fraction. Thus, we re-estimate our baseline specification on a smaller subset of schools in which students' prior achievement would need to change by less than a third of a quintile on average to match the statewide uniform distribution of quintiles. Figure **??** shows that the point estimates are roughly the same with the restricted sample, but with larger confidence intervals.

12 Conclusion

In this paper we use rich administrative data to identify the treatment effects of changing the honors track selectivity as functions of the share of students permitted to enroll in honors, with separate functions estimated for each quintile of an index of student preparedness. Importantly, our approach explicitly accommodates endogenous self-sorting of students into the honors and regular tracks conditional on the administrator-determined capacity of the honors track. We then show that our set of estimated treatment effect functions suffice to determine the optimal share of students in each track in an administrator's planning problem.

We obtain the result that the optimal share of students in the honors is between 20 and 30%. Based on results from our baseline specification, if all the schools switched from their current honors track sizes (including the absence of an honors program) to one with 20 to 30% of students in it, North Carolina public school student would gain over 0.02 SDs in test score performance on average. Altering the selectivity of the honors track thus represents a low cost method to improve test score performance, particularly for larger schools that are already offering the relevant courses in multiple class periods. Importantly, there is no tradeoff between efficiency and equity, as highly prepared and moderately prepared students benefit considerably from small honors tracks (between 0.04 and 0.07 SDs) relative to the

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

absence of tracking, while less prepared students only begin to experience losses when the honors track expands to nearly half the student population. Because these small per-student gains apply to such a wide population of students and high schools, our back-of-the-envelope calculations suggest that they could translate to aggregate skill development worth millions of dollars in future earnings potential. To provide reassurance about the validity of these findings, we show that they are extremely robust across several alternative specifications featuring different samples, functional form assumptions, or sources of variation that remove different sources of endogeneity.

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

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

In addition, the North Carolina context we consider provides strong incentives to keep the breadth of material covered by the course similar among both tracks in order to prepare all students for a common statewide standardized exam. This feature is essential for generating internally valid estimates by facilitating comparisons on a single achievement metric. However, we cannot verify external validity for contexts in which different tracks have substantially different curricula (e.g. Advanced Placement or International Baccalaureate), though we have no *a priori* reason to believe that our results would not generalize.

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