Academic Peer Effects with Different Group Assignment Policies: Residential Tracking versus Random Assignment*

Robert Garlick†

October 21, 2012

Abstract

A growing literature in the economics of education emphasizes the importance of peer effects in determining students' academic outcomes. Most of the existing empirical literature uses natural experiments in which students are randomly assigned to peer groups to identify the causal effect of peer characteristics. However, their results provide no direct evidence regarding the policy-relevant problem of which group assignment policies produce better or worse aggregate outcomes. I address this gap in the literature by contrasting the distribution of test scores under two common policies for assigning students to residential peer groups: tracking and random assignment.

I find that tracking reduces mean GPA by approximately 0.12 standard deviations (one quarter of the black-white GPA gap) relative to random assignment. This is driven by very large negative effects on the lower tail of the GPA distribution (up to 0.5 standard deviations) and near-zero, insignificant effects on the upper tail. These effects are robust to a variety of strategies to control for selection on observed and unobserved characteristics.

I explore whether these effects could have been predicted using random variation in dormitory composition generated under the random assignment policy. I estimate a flexible education production function and show that the parameters do not predict the magnitude of the treatment effect of tracking. This problem may reflect out of sample prediction problems or students' behavioral responses to changes in peer group composition, both of which are present in this setting.

JEL codes: I23, I24, I25, O15

*I thank Manuela Angelucci, Raj Arunachalam, Emily Beam, John Bound, Tanya Byker, John DiNardo, Susan Godlonton, Andrew Goodman-Bacon, Italo Gutierrez, Brad Hershbein, Brian Jacob, David Lam, Rebecca Thornton, Jeffrey Smith, Dean Yang; seminar participants at the University of Michigan, University of Cape Town, and UM-MSU-UWO Labor Day; and conference participants at CSAE, EconCon, ESSA, MEA, MIEDC, PACDEV, and SOLE for helpful suggestions. Jane Hendry, Charmaine January, and Josiah Mavundla provided invaluable assistance in obtaining the data used in this project. All errors are my own.

†University of Michigan; rgarlick@umich.edu

1 Introduction

Group structures are ubiquitous in education and group composition may have important effects on education outcomes. Students in different classrooms, living environments, schools, and social groups are exposed to different peer groups, receive different education inputs, and face differential institutional environments. A growing body of empirical evidence shows that students' peer groups influence their education outcomes even when resource and institutional differences across groups are negligible. Peer effects have been documented on students' college GPAs (Sacerdote, 2001), standardized test scores (Hoxby, 2000), college entrance examinations (Ding and Lehrer, 2007), study habits (Stinebrickner and Stinebrickner, 2006), major choices (di Giorgi, Pellizzari, and Redaelli, 2010), job search (Marmaros and Sacerdote, 2002), and social networks (Marmaros and Sacerdote, 2006). Academic peer effects play a role in both empirical and theoretical studies of classroom tracking (Arnott, 1987; Duflo, Dupas, and Kremer, 2011), neighborhood segregation (Benabou, 1996; Kling, Liebman, and Katz, 2007), school choice and vouchers (Epple and Romano, 1998; Hsieh and Urquiola, 2006), and school integration and busing policies (Angrist and Lang, 2004).[1] The majority of these studies focus on the effect of assignment to or selection into different peer groups for a given group assignment or selection process.

This paper asks a subtly different question: What are the effects of different policies for assigning students to groups? This contributes to a small but growing empirical literature on optimal group design. Comparison of different group assignment policies corresponds to a clear social planning problem: How should students be assigned to groups in order to maximize some measure of academic output, subject to a given distribution of student characteristics? Different group assignment policies leave the marginal distribution of inputs into the education production process unchanged. This raises the possibility of increasing academic output without the pecuniary costs associated with most education interventions. Low cost education interventions are particularly attractive for developing country education systems that often face serious resource shortages. I study the relative effects of two common policies for assigning college students to residential groups: random assignment and academic tracking. Under random assignment, each residential group is approximately representative of the entire population of students. Under tracking, each student is placed in a residential group of peers with similar baseline academic performance. I use a natural experiment at the University of Cape Town in South Africa to show that tracking substantially harms low-scoring students and has little effect on high-scoring students. The net effect is a decrease in mean academic output and an increase in inequality of academic output.

This cross-policy comparison is important because peer effects estimated under one assignment policy can only predict the effect of a change in the assignment policy under strong assumptions.

[1] Some evidence suggests that peer effects estimates are sensitive to measurement of peer groups and peer characteristics (Cooley, 2010; Foster, 2006; Stinebrickner and Stinebrickner, 2006). Empirical results that fail to find evidence of peer effects may constitute genuine evidence against peer effects or reflect these measurement challenges.

For example, randomly assigning students to dormitories allows researchers to identify the effect of residential peers' academic ability on students' academic outcomes. These "within-policy" results can predict how a student's outcome will change if the proportion of high ability students in her dormitory increases marginally, for example from 40% to 60%. However, there is a low probability that random assignment will generate "tracking-like" dormitories where the proportion of high ability students is near 0 or 100%.[2] Any attempt to predict the effect of tracking students into dormitories thus relies on out-of-sample extrapolation. Even if this out-of-sample extrapolation is valid, the policy change may produce general equilibrium responses that cannot be predicted using within-policy variation in group composition. For example, students may have a preference for academically diverse social groups. Their residential peer group satisfies this condition under the random assignment policy but not under the tracking policy. Students would then be more likely to form social ties outside their dormitory under the tracking policy, a behavior that peer effects estimated under random assignment could not predict.

Given the challenges of cross-policy prediction, the empirical and methodological literature on this topic is sparse. Bhattacharya (2009) and Graham, Imbens, and Ridder (2011) develop methods for predicting the effect of non-random assignment policies using peer effects estimated under random assignment. However, these approaches impose strong assumptions that rule out both the out-of-sample extrapolation and general equilibrium problems described above. Carrell, Sacerdote, and West (2012) and Duflo, Dupas, and Kremer (2011) provide what appears to be the only direct empirical evidence regarding the effect of different policies for creating peer groups. Both studies find that peer effects estimated under one assignment policy do not predict the result of changing that assignment policy.

A better understanding of how peer effects operate under different group assignment may help to understand important policy debates in education. In particular, an extensive empirical and theoretical literature has studied the effect of tracking students into classrooms or schools (Betts, 2011). Most empirical studies estimate reduced form effects of tracking relative to other assignment policies but cannot separate the mechanisms driving their results. These results may be driven by peer effects or by differences between tracked and untracked groups in curricula, instruction, or other resources. A small number of papers use careful data collection strategies to assess the relative importance of different mechanisms (Ding and Lehrer, 2007; Duflo, Dupas, and Kremer, 2011; Pop-Eleches and Urquiola, 2012) but the role of peer effects in academic tracking remains poorly understood. Group assignment or selection processes also play potentially important roles in debates around school choice (Epple and Romano, 1998; Hsieh and Urquiola, 2006) and school integration (Angrist and Lang, 2004).

This paper provides a bridge between the literatures on academic tracking and on optimal

[2] In contrast, random assignment to smaller groups such as roommate pairs may plausibly generate substantial numbers of "high-high" and "low-low" pairs. These pairs could in principle be used to predict the effect of changing the assignment policy to tracking. I thank Todd Stinebrickner for pointing out this distinction. The concern about general equilibrium responses still applies in this smaller group setting.

design of peer groups. I study a natural experiment where students were assigned to residential groups (dormitories) using tracking for several years and thereafter by random assignment. I contrast the distribution of dormitory students' academic outcomes under the two policies and show that mean academic performance was lower and inequality higher under tracking. I argue that differences between the tracked and untracked groups are small and show that the results are highly robust to accounting for these differences. I use non-dormitory students who live in private accommodation as a control group to remove any time trends, flexibly control for differences in students' observed characteristics and use university admission policies to construct instrumental variables to control for differences in students' unobservd baseline characteristics. The results are highly robust to these strategies and to several additional sensitivity analyses. My findings show that alternative peer group assignment policies can have substantial effects on academic performance. In particular, the distribution of outcomes under tracking is worse than under random assignment for most plausible social welfare criteria in this setting. The tracked and untracked groups do not experience different curricula, instruction or institutional environments. While there are differences in physical facilities across dormitories, these are stable through time and so do not differ between the tracked and untracked groups. The differences in academic performance must thus be driven by peer effects, suggesting that a pure peer effects mechanism is an important part of the effect of tracking in other settings.

This paper also explores the relationship between peer effects implied by the cross-policy results and peer effects estimated under a given policy. In particular, random student assignment to dormitories allows me to consistently estimate the effect of peers' baseline characteristics on students' academic outcomes. These estimates confirm that students are affected by their residential peers' academic proficiency, measured by high school graduation test results, and that low-scoring students are significantly more affected than high-scoring students. This pattern is qualitatively consistent with the results of the main cross-policy analysis but predicts far smaller effects of tracking than I observe. This may reflect out-of-sample prediction problems or behavioral responses by students. I show that the effect of peer characteristics on student outcomes is much stronger when those peers are of the student's own race group. This suggests that spatial proximity increases the probability that students will interact in a way that generates peer effects but that interaction also depends on students' own preferences and behavior. Behavioral responses may thus mediate peer effects on academic performance and explain the failure of within-policy estimates to predict the effect of policy changes.

This paper's substantive contribution is facilitated by several econometric innovations. Standard methods for comparing academic outcomes under two different policies typically focus on the mean, for all students or for particular subgroups of interest. This is informative but more can be learned by contrasting the full distributions of outcomes. I therefore estimate the counterfactual distribution of academic outcomes that tracked students would have achieved if tracking were not implemented. This employs the "nonlinear difference-in-differences" model proposed by Athey and Imbens (2006) and I provide one of the first empirical applications of this method. I also extend the model to

flexibly account for differences in the distribution of observed student characteristics such as race and gender. Having constructed the full counterfactual distribution of outcomes, I can estimate the effect of tracking on summary statistics of interest such as the level of academic inequality. I also note that tracking may affect students' academic mobility, measured by the probability that low-ranked students in high school will become high-ranked in university and vice versa. I develop the idea of a "mobility treatment effect" to quantify this effect and develop the appropriate statistical theory in a companion paper (Garlick, 2012).

These findings suggest that group assignment policies may have substantial effects on academic performance even when other education input are held constant. Such policy changes impose few pecuniary costs and so offer policymakers an attractive method for boosting academic performance. Different group assignment policies may impose transfers across different groups of students and, as academic outputs are not tradeable, Pareto ranking of different policies will often be impossible. In this setting, the large losses that tracking imposes on low-scoring students are not accompanied by significant positive effects on high-scoring students. Most social welfare criteria would thus prefer random assignment to tracking but care should be taken in applying this result in other settings. For example, the distribution of students' academic proficiency at the University of Cape Town is probably wider than in most tertiary institutions in developed countries. The university attracts many of the highest performing students in South Africa but also admits many academically promising students from low-performing high schools with weak preparation for further study.[3] These are the students most affected by tracking, so my results are particularly relevant to policy debates about how best to help academically underprepared students succeed in tertiary education (Roderick, Nagaoka, and Coca, 2009).

Organization of the paper: Section 2 discusses the setting that I study and the two assignment policies. Students were assigned to dormitories using a tracking policy up to the 2005 academic year and randomly assigned to dormitories from the 2006 academic year onward. I therefore employ a difference-in-differences design that compares the academic outcomes of dormitory students under the two assignment policies, using non-dormitory students to control for changes through time in student characteristics or university policies. I briefly discuss some potential threats to the validity of this design and show in the appendices that my results are highly robust to such concerns. This section also quantifies the change in the distribution of peer characteristics across groups, showing that the policy induced a large reallocation in peer quality.

Section 3, which compares students' outcomes across the tracking and random assignment policies, is the main empirical component of the paper. I find that tracking reduces mean GPA by approximately 0.12 standard deviations. This result is driven by a large negative effect on the students with low high school grades (up to 0.25 standard deviations) and a small and imprecisely estimated positive effect on the upper tail. Quantile treatment effects verify that tracking substantially harms

[3]These students may have relatively high levels of academic ability. However, they perform poorly in the standardized high school graduation tests that I use to measure academic proficiency.

the left tail of the GPA distribution and has almost no effect on the right tail. I show that this pattern of treatment effects implies substantial increases in standard measures of inequality, suggesting that tracking has negative efficiency and equity effects. However, I find no effect of tracking on mobility, measured the probability that students change their rank in the distribution of academic achievement from high school to university. This suggests that tracking disproportionately reduces low scoring students' level of GPA but leaves their rank largely unchanged.

Section 4 explores what could have been learned about peer effects without a policy change, if I had only observed students during the random assignment period. I exploit the fact that random assignment to dormitories balances peer group means only in expectation and estimate the effect of random variation in dormitory-level measures of student characteristics. I show that students' GPAs are affected by the characteristics of their residential peers and that students with low high school grades are more strongly affected. I use empirical and theoretical arguments to highlight the difficulty of predict the effects of tracking using these results.

Section 5 explores the economic behavior driving the negative effect of tracking. I begin by showing that the treatment effects occur early in the semester, not just during final examinations. I then evaluate three competing models of how spatial proximity due to dormitory assignment generates peer effects. I reject a "noisy neighbor" model in which spatial proximity alone is sufficient to generate peer effects. The data provide weak evidence against an "academic collaboration" model in which spatially proximate peers influence outcomes primarily by working on the same academic tasks. The clearest pattern in the data is that spatial peer effects operate entirely within race group. I interpret this as evidence in favor of a "buddy" model, in which peers influence student outcomes through time use and academic focus, rather than direct assistance with studying or problem sets.

Section 6 concludes and discusses the implications of these results. I argue that this paper provides important and novel evidence on the academic effects of residential tracking, which are driven entirely by differences in peer groups across dormitories. This finding is directly relevant for any policy maker who assigns student to residential groups. It also indicates that a pure peer effect mechanism may be important in other group assignment problems. For example, existing studies on academic tracking by classroom or school typically conflate differences in peer group characteristics with differences in instructor behavior or school resources. My results provide direct evidence regarding the importance of differences in peer characteristics. This result is also relevant to the literature on neighborhood segregation by family socioeconomic status (typically highly correlated with academic proficiency measures used for tracking).

The appendices present a variety of robustness checks to show that the results in section 3 are not driven by violations of the identifying assumption underlying the difference-in-differences design. This design assumes that the time change in students' academic outcomes between the tracking and randomization periods would have been equal for dormitory students (the treatment group) and non-dormitory students (the control group) if the policy change had not occurred. Appendix A.1 shows that

the results are not driven by changes in course-taking behavior between the two periods. Appendix A.2 shows that the results are not driven by students strategically selecting whether or not to live in a dormitory. I construct an instrument for treatment status using the university's admissions policies and show that the instrumental variables and least squares results are very similar. Appendix A.3 then presents a sensitivity analysis showing that an implausibly large change in students' unobserved characteristics would be needed to explain the observed treatment effects. Finally, appendix A.4 assumes that the treatment effects have been correctly estimated and shows that inferences regarding these effects are robust to alternative test procedures.

Appendix B provides more detail regarding the nonlinear difference-in-differences model employed in section 3. I present an intuitive explanation of the model's identification results and discuss how the identifying assumptions can be relaxed to accommodate differences in students' observed characteristics.

Related literature: This paper relates to two literatures in the economics of education: peer effects and tracking. I study the effect of tracking students into groups whose only substantial difference is their composition. Unlike most studies of tracking, groups do not face different instructors, curricula or resources. I therefore estimate a treatment effect of tracking arising directly and exclusively from a shift in peer group composition and so provide a bridge between the tracking literature and the recent peer effects literature on optimal group assignment. This section highlights some particularly relevant results from these literatures. For more comprehensive reviews of the literature, see Epple and Romano (2011) and Sacerdote (2011) on peer effects in education and Betts (2011) on academic tracking.

Manski (1993) provided the first formal treatment of the methodological challenges involved in estimating peer effects. He notes that in a linear regression of own academic outcomes on peers' baseline characteristics, a non-zero coefficient may arise for multiple reasons. The coefficient may reflect a causal effect of peer characteristics on own outcomes but may also arise due to correlated unobserved characteristics due to endogeneous group formation or correlated shocks because students are exposed to the same environment as their peers. Most empirical papers therefore study settings in which students' peer groups vary for reasons uncorrelated with their own unobserved characteristics and include higher-level fixed effects to control for correlated shocks. The most directly relevant set of papers study the effect of random assignment to roommates with different SAT scores while using dormitory-level fixed effects to address correlated shocks (Sacerdote, 2001; Stinebrickner and Stinebrickner, 2006; Zimmerman, 2003).[4] These studies typically find small to moderate effects of peer characteristics on academic outcomes. Cooley (2010) and Stinebrickner and Stinebrickner (2006)

[4]Other studies estimate the effect of cohort-level variation within school-by-grade units in gender composition, race composition or test scores (Hanushek, Kain, Markman, and Rivkin, 2008; Hoxby, 2000; Lavy and Schlosser, 2011) or the effect of approximately random assignment of students to classrooms (Ammermueller and Pischke, 2009; Vigdor and Nechyba, 2007; Kang, 207). Another literature uses natural experiments in which students' peer groups are changed by school integration policies (Angrist and Lang, 2004; Hoxby and Weingarth, 2006) or natural disasters (Imberman, Sacerdote, and Kugler, 2012).

emphasize that estimates of peer effects may be sensitive to how peer characteristics are measured. Hoxby and Weingarth (2006) and several more recent paper estimate that the functional form of reduced form estimating equations have significant implications for interpretation of the results. Foster (2006) and Carrell, Fullerton, and West (2009) note that administative units such as dormitories or classrooms may be poor proxies for students' true peer groups, which are typically unobserved to researchers.[5] I discuss the relevance of these considerations to this paper in sections 2, 4 and 5 respectively.

This paper is closely related to Carrell, Sacerdote, and West (2012), which examines the relative effects of two different policies for assigning freshmen to squadrons at the US Air Force Academy. They find that randomly assigning students leads to higher average GPAs than "mixing" students, so that some groups comprise only high and low ability students, while others comprise only students of moderate ability. However, using peer effects estimated across randomly created groups predicted that GPAs would be higher under the latter policy than the former. This result highlights the danger of predicting the effects of one policy using results estimated under another policy and underscores the value of papers that explicitly compare outcomes across different policies.

The tracking literature can be divided into two broad strands: the first studies the effect of tracking compared to some other group assignment policies and the second studies the effect of assignment to different tracks. Betts (2011) reviews the first literature, noting that tracked and untracked groups often differ in curriculum, instructor characteristics, and school resources. Few papers are able to disentangle the relative importance of these factors and so cannot determine whether differences in the distribution of peer characteristics across tracked and untracked groups lead to differences in outcomes. Duflo, Dupas, and Kremer (2011) provide an important exception: they use a field experiment to show that tracking Kenyan students into classrooms increases test scores for high and low track students relative to randomly assigning them. They also estimate that being randomly assigned to a classroom with higher scoring peers raises students' test scores. They reconcile these results by suggesting that students' in the low track classrooms gained more from focused instruction than they lost from weaker peer groups.

Papers in the second strand of the tracking literature typically estimate the effect of assignment to a high instead of low track. Tracks often differ along multiple dimensions – including peer characteristics, curriculum, instructor characteristics – so these effects do not directly test whether peers matter for academic outcomes. Some important exceptions include Ding and Lehrer (2007) and Pop-Eleches and Urquiola (2012) who study selective high schools in China and Romania respectively. The former paper finds large effects of attending selective schools and attributes much of this to measured teacher quality and a small portion to peers. The latter paper finds large effects of attending selective schools and selective tracks within schools but note that qualification for higher tracks is associated with substantial changes in instructor, parent, and student behavior.

[5] Guryan, Jacob, Klopfer, and Groff (2008) is one the few papers that directly observes student interactions directly instead of proxying them with group assignment.

This paper's primary contribution is to the empirical literature but it also builds off a rich theoretical literature relating to peer effects and tracking. The theoretical discussion in section 4 builds off models of peer interaction discussed by Benabou (1996) and Epple and Romano (2011). Blume, Brock, Durlauf, and Ioannides (2011) and Brock and Durlauf (2011) review the literature on social interactions models. They place particular emphasis on models that attach a theoretical interpretation to the common estimands in the empirical literature. Graham (2011) reviews the literature on econometric methods for identification and estimation of such models.

The paper's empirical methodology builds off a number of recent results in the heterogeneous treatment effects literature. I estimate quantile treatment effects of tracking using a nonlinear difference-in-differences model that recovers the full counterfactual distribution of GPAs that the tracked students would have obtained if they had instead been assigned to dormitories using random assignment (Athey and Imbens, 2006). I also implement an extension to the original Athey-Imbens models that incorporates changes in the distribution of observed student characteristics and discuss this in detail in appendix B. This extension closely follows the reweighting methods proposed by DiNardo, Fortin, and Lemiuex (1996) and formalized by Hahn (1998) and Firpo (2007), amongst others. I am able to use this counterfactual distribution of GPAs to estimate the effect of tracking on inequality, following Firpo (2010). Finallly, the treatment effects of tracking on academic mobility that I present in section 3 are based on identification and estimation results developed in a companion paper (Garlick, 2012).

2 Experimental Setting and Research Design

Experimental setting: The policy experiment I study was carried out at the University of Cape Town (UCT) in South Africa. UCT is a selective research university whose student body is not representative of South Africa's population. However, the student body is substantially more heterogeneous than in US universities in which much of the prior peer effects literature is located. The university admits between 3500 and 4000 incoming students each year. Approximately half of these students live in private accommodation and half live in the 16 university dormitories, each of which accommodates between 35 and 237 students.[6] Incoming students before the 2006 academic year were tracked into dormitories based on their performance in a standardized high school graduation exam, with high-scoring students living in different dormitories to low-scoring students. From 2006 onward, students were assigned to dormitories by a random number generator. Under both policies, dormitory assignment did not directly affect the students' courses or instruction. Students from all dormitories and non-dormitory students were taught the same material in the same classes by the same instructors,

[6]The mean dormitory size over the study period is 123 students and the 10^{th}, 50^{th} and 90^{th} percentiles are 50, 112 and 216 students. There are no substantial changes in dormitory size between the two periods. Fourteen dormitories were open under both policies, one was open only under the tracking policy and one was open only under the random assignment policy. My results are robust to using only the fourteen dormitories open under both policies.

so the dormitory assignment effect operated solely through students' living environment and spatial peer group. There is some variation in dormitories' physical facilities and proximity to campus and to leisure opportunities but controling for these differences using dormitory fixed effects makes little difference to my results.

The policy change: Under the tracking policy, students were assigned to dormitories based on their result in a standardized high school graduation examination, which was also used for admissions decisions.[7] There were, however, two deviations from a pure tracking assignment, in which dormitories partition the distribution of high school grades. First, the assignment regime included an affirmative action component, so black students required lower scores for assignment to high track dormitories than white students.[8] Second, students who applied late to the university were waitlisted for places in the dormitories and assigned when admitted students declined their admission offer or deregistered. The student at the top of the waitlist was assigned to the first space that became available, without regard to their high school grades. These exceptions increase within-dormitory variation and decrease between-dormitory variation relative to pure tracking. This may attenuate the estimated treatment effect of tracking. Furthermore, the race-specific thresholds and *ad hoc* assignment of late applicants mean that there is non-trivial overlap in student's high school grades across dormitories. There is thus insufficient mass near the dormitory-specific thresholds to permit estimating the effect of assignment to one dormitory or another by a regression discontinuity design.

Under the mixing policy, students were randomly assigned to dormitories, conditional on their race. At the time of their admission, students were assigned to a dormitory based on a race-blind random number generator. The staff member responsible for assignments regularly checked the racial composition of each dormitory and if he believed that one race group was underrepresented, he manually assigned the next few applicants of the underrepresented race to that dormitory. The criteria for determining "underrepresentation" were left to the discretion of the staff member, who reported making "only a few" reassignments each year.

This raises the possible concern that students were able to manipulate their dormitory assignment by lobbying staff members involved in the assignment process. When interviewed, the director of the admissions office acknowledged that this was a risk under both the tracking and random assigment regimes but stated that "everyone involved in the process was instructed to present a united front ... that residence assignments were final." Informal discussion with both students and staff suggest that this policy was strictly enforced. Such manipulation, if it took place, complicates interpretation of the empirical results but does not affect the internal validity of my research design.

Under both the tracking and random assignment policies, 11 of the 16 dormitories were single-

[7]The university's admissions office converts international students' grades in A-level, International Baccalaureate, and Higher International General Certificate in Secondary Education examinations into equivalent grades on the South African graduation examination.

[8]South Africa's population was divided into four groups under *apartheid*'s racial classification system: black, white, Indian and coloured. Given the ongoing salience of racial divisions, these distinctions are still in widespread use in social science research, public discourse, and government policy.

sex. This meant that in the tracking period, male and female students with the same high school grades in general lived in different dormitories. Under both policies, students lived in dormitories for at most two years before moving into private accommodation or university-owned apartments off campus. Hence, in 2006, the dormitory system contained both randomly assigned first year students and tracked second year students, who continued to live in the dormitory to which they were originally assigned. This group of first year students were not, therefore, exposed to the same mixing policy as those in 2007 and 2008 and so I omit 2006 from my main sample.[9]

Quantifying the extent of the policy change: To demonstrate that the change in dormitory assignment policy had a substantial effect on the distribution of student characteristics across dormitories, figure 1 shows the mean and standard deviation of high school grades in each dormitory-year unit.[10] The tracking policy more than doubled the range of dormitory means from $(-0.2, 0.6)$ to $(-0.8, 12)$. The within-dormitory correlation coefficient rises from 0.12 under random assignment to 0.36 under tracking, which also emphasizes the substantial effect of the policy change on the distribution of students across dormitories. The within-dormitory correlation coefficient under tracking is still substantially less than one because of the *ad hoc* assignment of waitlisted applicants, race-specific assignment thresholds, and long lower tail of the grade distribution (clearly visible in figure 3). This long tail creates a relatively high within-dormitory variance in "low-track" dormitories.

The figure also highlights some of the departures from pure tracking. First, the dormitory means under tracking are often relatively close together, because there are often "paired" male and female dormitories with similar scores. Second, the four observations in the left of the figure are clear outliers under both assignment policies. This is the university's only "self-catering" dormitory, in which students purchase and prepare their own food rather than eating in communal dining halls. Students are permitted to request assignment to this dormitory and pay approximately 40% lower fees than in other dormitories. In practice it contains disproportionately many black students (who are likely to come from poorer households) and students with low high school grades. My results are entirely robust to excluding students in this dormitory, who make up only 1.4% of the sample.

Figure 2 provides an alternative depiction of the effect of the policy on the distribution of students across dormitories. This figure shows the difference in mean peer group high school grades between the tracking and random assignment periods, for students at each percentile of the high school grade distribution. Students in the bottom decile lived with peers whose average high school grades were 0.5 standard deviations lower under tracking than mixing. Students in the top decile lived with peers whose average high school grades were 0.4 standard deviations lower under tracking than mixing. This provides a simple measure of the change in peer group ability (proxied by high school grades) for different groups of students. The fact that observationally similar students had very different peer groups under the two policies provides the identifying variation I use to estimate the treatment effect

[9] My results are robust to including this year in the random assignment period.
[10] I standardize grades to have a mean of zero and variance of one within each year.

of tracking.[11]

Research design: My basic empirical strategy uses a linear difference-in-differences model to compare the GPA of dormitory students under the tracking policy and the random assignment policy, with non-dormitory students employed as a control group to remove any time trends in student GPAs. This strategy identifies the average treatment effect on the treated (ATT) of one policy relative to the other, assuming that the time trends in student GPAs for the treatment group (dormitory students) and control group (non-dormitory students) would have been identical in the absence of a policy change. I treat random assignment as the default policy and so define the parameter of interest as the average treatment effect of tracking on the treated students:

$$\begin{aligned}\Delta^{ATT} &= \mathbb{E}\left[GPA(1)|D=1,T=1,X\right] - \mathbb{E}\left[GPA(0)|D=1,T=1,X\right] \\ &= \mathbb{E}\left[GPA(1)|D=1,T=1,X\right] - \mathbb{E}\left[GPA(0)|D=1,T=0,X\right] \\ &\quad - \big(\mathbb{E}\left[GPA(0)|D=0,T=1,X\right] - \mathbb{E}\left[GPA(0)|D=0,T=0,X\right]\big),\end{aligned} \tag{1}$$

where $GPA(1)$ and $GPA(0)$ respectively denote GPA for students who are and are not tracked, $D=1$ for dormitory students, $T=1$ for the period in which the tracking policy was in place, and X is a vector of students' demographic characteristics and high school grades. The second equality follows from the assumption of identical trends. I could alternatively define tracking as the default policy and estimate the treatment effect of random assignment - the results are numerically invariant to this choice. Note that difference-in-differences designs estimate "treatment on the treated" parameters, which are valid only for the treatment group (dormitory students) and do not necessarily identify the effect that treatment would have on the control group (non-dormitory students).

I estimate Δ^{ATT} in the following regression model

$$GPA = \alpha + \beta D + \gamma T + \Delta TD + f(X) + \epsilon \tag{2}$$

where $f(\cdot)$ is a flexible function of the covariates. In practice, my results are robust to a wide range of different specifications. Assuming that $f(\cdot)$ is correctly specified, $\hat{\Delta}$ in equation (2) is a consistent estimator of Δ^{ATT} provided the mean change in unobserved determinants of GPA from the random assignment to the tracking period is identical for dormitory and non-dormitory students:

$$\mathbb{E}\left[\epsilon|D=1,T=1,X\right] - \mathbb{E}\left[\epsilon|D=1,T=0,X\right] = \mathbb{E}\left[\epsilon|D=0,T=1,X\right] - \mathbb{E}\left[\epsilon|D=0,T=0,X\right]. \tag{3}$$

This condition is not directly testable but I argue for four reasons that it is a plausible assumption.

[11] I create the figure in four steps. First, I calculate the mean high school grade in each dormitory and assign this to each student as a measure of their peers' baseline ability performance. Second, I estimate a nonparametric regression of peers' baseline performance against own high school grades, separately for the tracking and random assignment period. I use a local linear regression with an Epanechnikov kernel and a plug-in bandwidth following Fan and Gijbels (1996). I allow the bandwidth to differ for the tracking and random assignment periods. Third, I calculate the difference between the two fitted curves at each percentile of the distribution of high school grades. Finally, I construct a 95% confidence interval from the 2.5 and 97.5 percentiles of a nonparametric bootstrap, stratifying by period.

First, table 1 shows the time trend in dormitory and non-dormitory students' observed characteristics between the tracking and random assignment periods. The first panel shows that the time trends in mean high school grades were equal for dormitory and non-dormitory students. The same condition holds for the proportion of students scoring mostly As and mostly Ds in their high school graduation examinations. This rules out one *a priori* plausible violation of assumption (3): that dormitory system might have attracted more high achieving students in the tracking period than the random assignment period. The second panel shows that the time trends in the race, gender, and nationality composition of the student body were all approximately equal. The only characteristic that clearly violates assumption (3) is language, as the proportion of English–speaking students in the dormitories is higher in the tracking than the random assignment period.

Panel C shows that the probability of starting at the university in the random assignment period (2007 or 2008) after finishing high school early enough (2005 or earlier) to start during the tracking period is equal for dormitory and non-dormitory students. This suggests that students did not strategically delay their entrance to the university in order to manipulate the policy under which they were assigned. There is also no sign of such selection when restricting the test only to students near the top or bottom of the distribution of high school grades. These results show that the assumption of equal trends amongst treatment and control groups holds for students' observed characteristics and provides reassurance that the assumption may also hold for students unobserved characteristics. I explore the consequences of violations of this assumption in appendix A.3.

Second, neither assignment policy was well-publicized and little public information was made available about the policy change. The assignment policy in place at the time was not stated in the application and promotion materials distributed to potential students. Individual campus visits and other recruitment strategies employed in the US are rare, so most students' only interaction with the university was through promotional materials, presentations by admissions staff at their high school, or informal interaction with current students. The assignment policy was available on the university's website, but this was one brief part of a lengthy document containing all information about student housing. The change in policy was not announced in university or external media. Informal discussion with students suggested that relatively few were aware of the assignment policies when they started their studies and those who were aware tended to have older siblings or friends at the university.

Third, admissions staff report that students typically live in dormitories if and only if they live outside Cape Town. The dormitory system is somewhat oversubscribed and so at most 4% of dormitory places are available for students from Cape Town and its suburbs. Other local students tend to live with their families, as the university is within one hour drive of most parts of the greater Cape Town area. Students from outside Cape Town could, in principle, rent private accommodation rather than live in dormitories. However, private accommodation is typically considerably more expensive: a single room and three meals a day in a university dormitory in 2010 cost approximately 85% of the rent for

a single room in a shared apartment equivalently close to the university.[12] This suggests that students have relatively little discretion over whether they lived in a dormitory or in private accommodation. Panel D of table 1 verifies this fact: the proportion of students from schools inside and outside Cape Town who live in the dormitory system do not experience differential time trends. These equal time trends also hold for students at the top and bottom of the high school grade distribution. This again suggests that students' decision to enter the dormitory system was restricted by their residence, rather than strategic.[13]

Fourth, there were no simultaneous policy changes at the university that are likely to affect the composition of the student body or their performance. The only change to the administration of the dormitory system was the closure of one old dormitory in 2006 and the opening of a new dormitory in 2007. My results are robust to omitting students in those dormitories. The admissions and financial aid policies at the university stayed constant over this period. The fee policy did change between 2005 and 2006: students were previously charged a flat rate for enrollment and this was replaced by a fee–per-credit hour system. I show in appendix A.1 that the number of courses students took did not change over this period, reflecting the fact that the first year curriculum at requires a fairly standard number of courses. Inspection of minutes of Senate and other academic committee meetings shows no evidence of changes in grading standards or criteria at this time.

This discussion suggests that the "equal trends" assumption required for validity of the difference-in-differences design is plausible. In particular, the institutional setting allows limited scope for students to select strategically whether to live in dormitories in responses to the policy change: the change was not widely advertised and students' home location largely determined their admission to the dormitory system. Consistent with this claim, the time trends in observed characteristics for dormitory and non-dormitory students are approximately equal. Finally, there is little evidence of substantial changes in other policies at the university over this time period. The appendices also present detailed robustness checks that demonstrate that my results are robust to failures of the identifying assumption.

3 Treatment Effects of Tracking

I begin by estimating the average treatment effect of tracking on the treated students. The direction of this effect is not *a priori* clear: if tracking raises strong students' GPAs and lowers those of weak students, the sign of the average effect is ambiguous. I show that the data point to a large negative effect of tracking and then go on to explore how this effect varies throughout the distribution of student performance.

[12] The apartment rental rate is the average of the first ten apartments listed on the website `www.gumtree.co.za` on 30 October 2011 in the same neighborhood as the university.

[13] There are some students from Cape Town high schools who do live in dormitories and *vice versa*. This arises because admissions decisions are based on home location and this is imperfectly proxied by school location: some students at Cape Town schools live outside Cape Town and *vice versa*. Anecdotally, these students report either commuting long distances or living in hostels or dormitories during high school.

Average treatment effect of tracking: Table 2 reports estimates of the average treatment effect of tracking from the difference-in-differences model in equation (2). Column (1) reports a treatment effect of -0.13 standard deviations of GPA without conditioning on students' demographic characteristics or high school grades.[14] Standard errors are estimated using a cluster-robust variance matrix estimator that allows unrestricted correlation in students' unobserved characteristics at the dormitory-year level and the point estimate is significant at the 10% level. Adding dormitory fixed effects to remove any time-invariant characteristics of the dormitories that might affect students' GPAs (such as access to study facilities, proximity to campus, shared rooms) reduces the treatment effect to -0.11 standard deviations but reduces the standard error by a larger margin. Column (3) adds year fixed effects to remove year-specific GPA shocks that affect both the treatment and control groups and a flexible set of conditioning variables: student gender, language, nationality, race, a cubic in high school grades and all possible two- and three-way interactions between these variables. Column (4) conditions on dormitory fixed effects, year fixed effects, and student characteristics, yielding a precisely estimated treatment effect of -0.12 standard deviations. Taken together, these results suggest that tracking reduced dormitory students' GPAs by approximately .12 standard deviations with a 95% confidence interval from .06 to .18 standard deviations, and that this effect cannot be explained by baseline characteristics of the students or dormitories.

This effect is equal to approximately 150% of the male-female GPA gap and 30% of the black-white gap. Given the salience of race in the South African context, this points to a large and economically meaningful treatment effect. Benchmarking the magnitude of this effect relative to the existing literature is complicated by the fact that there are so few prior studies of noninstructional tracking. The two most similar studies in the existing literature yield comparably sized treatment effects, assuming that standard deviation-based scales are commensurable across such different settings. First, Duflo, Dupas, and Kremer (2011) find that instructional tracking in Kenyan primary schools increases students' test scores by 0.14 to 0.18 standard deviations relative to random assignment. Second, Carrell, Sacerdote, and West (2012) find that their "optimal" rules for assigning students to noninstructional peer groups at the US Air Force Academy reduce students' GPAs by 0.08 to 0.1 standard deviations relative to random assignment. These effects are, however, considerably smaller than most of the resource- and instruction-based interventions in developing countries reviewed in Glewwe and Kremer (2006).

How does tracking affect different subgroups? This section disaggregates the negative average treatment effect of tracking using a variety of heterogeneous treatment effect estimators. These estimates consistently show large negative treatment effects near the bottom of the GPA distribution and approximately zero treatment effects on students near the top of the GPA distribution.

Table 3 reports average treatment effects on the treated for several demographic subgroups. The treatment effects on male and female students are equal, but black students experience a substantially

[14] I standardize GPA by subtracting the control group mean and dividing by the control group standard deviation in each year.

larger negative effect than white students (0.23 standard deviations compared to 0.14 standard deviations). This difference is not significant but it suggests that black students, who on average have lower high school grades and lower socio-economic status, are disproportionately affected by the tracking policy. There is a positive but entirely insignificant treatment effect on students of other race groups; the reason for this effect is unclear.

This evidence is consistent with the hypothesis that the negative effects of tracking are concentrated amongst lower performing students. Table 4 provides more direct support by estimating average treatment effects on the treated for each quartile of the distribution of high school grades. The effects are in fact negative in all four quartiles but are considerably larger below the median (0.22 to 0.26 standard deviations) than in the top quartile (0.09 standard deviations). These results clearly point to negative effects on weaker students but suggest that even those students with high baseline performance might have lower GPAs under tracking.

It is, however, possible that quartiles are too coarse a division and that there are some students in the top quartile who do benefit from tracking. I explore this possibility by performing a local linear regression of GPA on high school grades, separately for dormitory students in the tracking period, dormitory students in the random assignment period, non-dormitory students in the tracking period, and non-dormitory students in the random assignment period. I then use the local linear estimates $\hat{GPA}_{DT}(HS)$ for group D in period T to compute the difference-in-differences estimator

$$\hat{\Delta}^{ATT}(HS) = \hat{GPA}_{11}(HS) - \hat{GPA}_{10}(HS) - \hat{GPA}_{01}(HS) + \hat{GPA}_{00}(HS)$$

at each percentile of high school grades (HS). This provides a flexible test of whether tracking has a positive effect for any subset of students. Panel A of figure 4 plots $\hat{\Delta}^{ATT}(HS)$ and shows that the point estimates are negative for approximately 80% of the distribution of high school grades.[15] The treatment effect is equal to approximately one quarter of a standard deviation in much of the bottom quartile, while the positive effect in the upper quintile never exceeds one tenth of a standard deviation. The bootstrap confidence intervals are relatively wide but the estimates are significant from approximately the 15^{th} to the 65^{th} percentiles.[16] This nonparametric regression does not control for student or dormitory characteristics but results are robust to splitting the sample and estimating $\hat{\tau}^{ATT}(HS)$ separately by decile with the same conditioning variables used in the previous section.[17]

[15] I use an Epanechnikov kernel and a plug-in bandwidth chosen separately for each of the four nonparametric regressions (Fan and Gijbels, 1996). The results are robust to substantial changes in the bandwidth parameter.

[16] This bootstrap algorithm resamples dormitory-year clusters, computes the local linear estimate for the resampled data, orders the resultant point estimates from 1000 replications, and computes the confidence intervals as the difference between the 25^{th} and 975^{th} estimates. The ordering is implemented separately for each percentile of the distribution of high school grades so the confidence intervals should be interpreted as providing only pointwise, not uniform, coverage. These confidence intervals depend on non-pivotal statistics and so do not offer asymptotic refinement but do avoid the need to rely on critical values from the standard normal distribution that may be a poor approximation to the finite sample distribution of the test statistic (Horowitz, 2001).

[17] Including controls in the local linear model is in principle possible. However, existing estimators for this "partially linear" model either assume that the distribution of high school points is strictly continuous (Yatchew, 1997) or require the choice of multiple bandwidth parameters and are highly sensitive to these choices (Robinson, 1988).

Quantile treatment effects of tracking: The nonlinear difference-in-differences model proposed by Athey and Imbens (2006) provides an alternative way to explore heterogeneous treatment effects, which I discuss in detail in appendix B. In brief, the model constructs the full counterfactual distribution (CDF) of GPAs for dormitory students under tracking *if tracking had not been implemented*, whereas the standard linear difference-in-differences model constructs only the counterfactual mean of this distribution. The observed and counterfactual distributions of grades are shown in the first panel of figure 5. The horizontal difference between these two distributions at each quantile is defined as the *quantile treatment effect on the treated* and these are plotted for each quantile in the second panel of figure 5. These are large and negative in the lower tail of the distribution, with effects of over 0.6 standard deviations at the 5^{th} percentile and more than 0.3 standard deviations at the 10^{th}. The effects are smaller at higher percentiles but remain negative for approximately 90% of the distribution. The estimates are relatively imprecisely estimated and statistically differ from zero only in the lowest quartile of the distribution[18] but the results clearly reaffirm that tracking has a significant negative effect on the lower tail of the distribution and little or no effect on the upper tail.[19]

The additional flexibility of the nonlinear difference-in-differences model comes at the cost of a stronger identifying assumption: that the distribution of baseline characteristics is constant through time for each group (dormitory and non-dormitory students). I show in appendix B that this restriction can be relaxed to allow for changes in observed student characteristics through time. This relaxation operates by reweighting students based on their baseline characteristics to equalize the distribution of observed baseline characteristics through time, following DiNardo, Fortin, and Lemiuex (1996) and Hirano, Imbens, and Ridder (2003). The first panel of figure 6 shows the quantile treatment effects estimated by this reweighted nonlinear difference-in-differences estimator. The results are exceptionally similar to those from the unadjusted estimator.

Figures 4 and 6 convey complementary but subtly different information about the nature of the treatment effects. The former figure presents *treatment effects for students at different points in the distribution of high school grades*. This shows that students with low high school grades were significantly harmed by the policy and that students with high grades may have been helped slightly. The latter figure presents *treatment effects at different points in the distribution of university GPA*. This shows that the lower tail of the GPA distribution dropped by a very large margin under tracking and the upper tail was unaffected. These latter statements are about the distribution of outcomes and make no claim about which students are at which point of this distribution. The former statements apply to specific subgroups of students defined by their high school grades. The two analyses present

[18] I again construct the confidence intervals using the percentile cluster bootstrap algorithm discussed in the previous footnote. The validity of the bootstrap for the nonlinear difference-in-differences estimator has not been formally established. However, Athey and Imbens (2006) report that tests based on bootstrap confidence intervals have coverage probabilities closer to their nominal levels than confidence intervals based on the analytical variance estimator that they derive.

[19] Note that these are treatment effects on quantiles of the outcome distribution, not treatment effects on individual students. The two concepts are equivalent only if tracking is a rank-preserving treatment, so that the students in the bottom of the observed distribution would also be at the bottom of the counterfactual distribution (Heckman, Smith, and Clements, 1997).

a consistent picture – tracking hurts weaker students and does little to help strong students – but do so using different methods.

Having constructed the full counterfactual distribution of GPAs, I can also compare a range of summary statistics for the observed and counterfactual distributions. In particular, the first row of table 5 shows that the average treatment effect on the treated from the nonlinear model is -0.13 standard deviations, or -0.1 standard deviations after conditioning on student characteristics and dormitory fixed effects. This is almost identical to the treatment effect estimated by the linear difference-in-differences model, which provides some reassurance that my results are not driven by differences in the two models' assumptions.

Treatment effects of tracking on inequality: The next four rows of table 5 show that standard measures of inequality are sharply higher under tracking. The interdecile range rose by approximately 15%, the interquartile range approximately 10%, the Gini coefficient by approximately 20% and the coefficient of variation by a considerably smaller margin.[20] These differences are all significant and suggest that any social welfare function that values both average GPA and equality of GPA would find tracking a particularly unattractive policy in this setting.[21]

Treatment effects of tracking on academic mobility: This section examines the relative effects of tracking and random assignment to dormitories on changes in students' rank mobility between high school and university. Specifically, I test whether the two policies have different effects on the probability that students will change their rank in the distribution of grades from high school to university. This complements the analyses in the previous section, which showed that tracking had a significant negative effect on GPAs in the lower tail of the distribution. This *level effect* demonstrates that random assignment helps low-achieving students to "catch up" to their peers. The *rank analysis* I present in this section demonstrates that random assignment is not enought to facilitate "overtaking" or rank changes between high school and university.

I begin by constructing transition matrices for the tracking and random assignment periods. Each p_{ij} element of these four-by-four matrices indicates the probability that an individual in quartile i of the distribution of high school grades will move to quartile j of the distribution of university grades. A diagonal matrix corresponds to zero mobility (every student remains in the same quartile), while $p_{ij} = 0.25 \; \forall i, j$ corresponds to complete mobility (students' final quartiles are unrelated to their initial quartile). I pool dormitory and non-dormitory students' grades to compute the quartiles but present the transition probabilities for only dormitory students.

The two transition matrices are reported in table 6. The first row of panel A shows that dormi-

[20] The Gini coefficient is only defined for random variables with strictly positive support. I therefore add 4 to the standardized GPA measure to ensure that all values are positive. This also ensures that the mean is non-zero, which necessary for estimating the coefficient of variation.

[21] Note that I construct the counterfactual distributions quantile-by-quantile so the interdecile and interquartile ranges are direct by-products of this process but the mean, Gini coefficient, and coefficient of variation require linear approximation between the quantiles. This linear approximation introduces non-classical measurement error into the estimation so the estimates should be interpreted with a degree of caution. I attempt to minimize the extent of the error by estimating the counterfactual distribution at 199 quantiles.

tory students in the tracking period who are in the bottom quartile of the high school grade distribution remain in the bottom quartile in university with probability 0.37 and move to the top quartile with probability 0.28. A visual inspection of panel A (tracking) and panel B (random assignment) suggests that the policy change had little effect on the transition probabilities.

Table 7 presents several summary measures of mobility in each period. I define a *mobility treatment effect* as the change in a summary mobility measure from the random assignment to the tracking period. The first row shows that the average probability that a student will move from one quartile to another is slightly higher than 0.8 in each period (Bartholomew, 1982). The average number of quartiles changes is approximately 0.9 in each period (Bartholomew, 1982) and the correlation between initial and final quartiles (measured by the second largest eigenvalue of each transition matrix) is approximately 0.55 (Sommers and Conlisk, 1979). None of the mobility treatment effects are significant and the magnitudes are very small, suggesting that the policy change raised the level of low-achieving students' GPAs but had little effect on their probability of overtaking students with higher grades in high school.

4 What Can Be Learned from Cross-Dormitory Variation?

The preceding section presents indirect evidence of the importance of peer effects in determining students' GPAs. My argument can be characterized as follows: students' GPAs changed sharply when the residential tracking policy was introduced, this change cannot be explained by differences in student or dormitory characteristics, no other simultaneous policy changes occurred, so the change must be due to the different peer groups created by the tracking regime. This section complements the argument by presenting a direct test for peer effects using cross-dormitory variation in peer characteristics.

Begin by considering a simple pseudo-structural model of students' GPAs adapted from Manski (1993) and Sacerdote (2001):

$$GPA_{ig} = \alpha_0 + \alpha_1 HS_{ig} + \alpha_2 \overline{HS}_g + \alpha_3 \overline{GPA}_g + \epsilon_{ig}, \tag{4}$$

where GPA_{ig} and HS_{ig} are the university GPA and high school grade respectively of student i in dormitory g. Define $\overline{HS}_g$ and $\overline{GPA}_g$ as the average GPA and high school grade of all students in dormitory g. In the language of Manski (1993), α_2 is an "exogeneous peer effect," in which students' GPAs are influenced by the baseline characteristics of their peers and α_3 is an "endogenous peer effect," representing a feedback loop between each students' GPA and that of her peers. It is clearly impossible to estimate equation (4) consistently, as the same variable appears on both sides of the equation. However, it is possible to evaluate equation (4) at the dormitory average, solve for

$$\overline{GPA}_g = \frac{\alpha_0}{1-\alpha_3} + \frac{\alpha_1 + \alpha_2}{1-\alpha_3}\overline{HS}_g + \frac{1}{1-\alpha_3}\bar{\epsilon}_g$$

and substitute this back into the pseudo-structural model to obtain the reduced form

$$\begin{aligned} GPA_{ig} &= \frac{\alpha_0}{1-\alpha_3} + \beta HS_{ig} + \frac{\alpha_2 + \alpha_1\alpha_3}{1-\alpha_3}\overline{HS}_g + \epsilon_{ig} + \frac{\alpha_3}{1-\alpha_3}\bar{\epsilon}_g \\ &\equiv \pi_0 + \pi_1 HS_{ig} + \pi_2 \overline{HS}_g + \eta_{ig}. \end{aligned} \tag{5}$$

If students are randomly assigned to groups, $\overline{HS}_g$ is uncorrelated with η_i, which is a student-specific deviation from the average dormitory unobserved characteristics. Then this model can be consistently estimated and $\pi_2 \neq 0$ if and only if some peer effect exists: $\alpha_2 \neq 0$ or $\alpha_3 \neq 0$. (Unless one peer effect is negative, a possibility seldom considered in the literature.) Sacerdote (2001) therefore proposes $H_0 : \hat{\pi}_2 = 0$ as a reduced-form test for the existence of either peer effect.

Table 9 reports the results of this test for the sample of all dormitory students in the mixing period, who are randomly assigned to dormitories. The value of $\hat{\pi}_2 \approx .24$ implies that a one standard deviation increase in peers' high school grades is associated with a one quarter standard deviation increase in each students' GPA in their first year of university. This result is robust to conditioning on students' demographic characteristics (column 2), to including dormitory fixed effects (column 3), and to excluding the one outlying dormitory discussed in section 2 (columns 4–6).

While it is impossible to recover the values of the structural parameters α_2 and α_3 from the reduced form coefficients π_1 and π_2, it is worth noting that if $\alpha_3 = 0$, then $\hat{\alpha}_2 = \hat{\pi}_2 \approx 0.24$ (standard error 0.09) and if $\alpha_2 = 0$, then $\hat{\alpha}_3 = \frac{\hat{\pi}_2}{\hat{\pi}_1 + \hat{\pi}_2} \approx 0.4$ (standard error 0.06). The latter estimate of the endogeneous peer effect or social multiplier implies that approximately two fifths of the gain from a dormitory-level increase in GPA spills over onto each student.

Note that the variation in $\overline{HS}_g$ used to estimate $\hat{\pi}_2$ arises because randomization balances dormitory means only in expectation, not in any particular finite sample. This means that the regressor of interest takes on only 30 values, which are relatively close together. This narrow support of the regressor and the cluster-robust variance estimator I use yield relatively large standard errors. This limited support also explains the large value of $\hat{\pi}_2$ relative to much of the existing peer effects literature. The point estimate suggests that a one standard deviation increase in peers' mean grades in high school increase own GPA by one quarter of a standard deviation. However, the range of $\overline{HS}_g$ is only .85 (after excluding the outlying dormitory), so moving from the "worst" to the "best" observed peer group would raise GPA by one fifth of a standard deviation.

The linear in means model is a standard point of departure in the empirical peer effects literature but it embodies a number of important limitations. In particular, the model requires that own and mean peer group grades are additively separable in the production function, so any reallocation of students between peer groups will leave the average GPA unchanged. This implies a zero average treatment effect of tracking, as well as any other reallocation policy. Combining this observation with the empirical results above suggests that the linear in means model is seriously misspecified.

I therefore estimate a more general model of the form

$$GPA_{ig} = f\left(HS_{ig}, \overline{HS}_g\right) + X_{ig}\Gamma + \eta_{ig} \tag{6}$$

where $f(\cdot,\cdot)$ is a polynomial function and X_{ig} is a vector of student characteristics and dormitory fixed effects. The polynomial function can be interpreted as a semiparametric series estimator or simply a flexible functional form. Cross validation suggests that the optimal order of the polynomial is 2,[22] so I estimate the model

$$GPA_{ig} = \beta_0 + \beta_1 HS_{ig} + \beta_{11} HS_{ig}^2 + \beta_2 \overline{HS}_g + \beta_{22} \overline{HS}_g^2 + \beta_{12} HS_{ig}\overline{HS}_g + X_{ig}\Gamma + \eta_{ig} \tag{7}$$

with and without student and without X_{ig}.

Table 10 reports estimates of the parameters of the augmented model (7). The key results are that $\hat{\beta}_{12}$ is consistently negative across all specifications, while the sign and magnitude of $\hat{\beta}_{22}$ are somewhat sensitive to specification of the control vector. The negative value of $\hat{\beta}_{12}$ indicates that students with low high school grades benefit more from an increase in mean peer high school grades than do students with higher grades. This implies that tracking, by blocking interaction between low and high scoring students, will hurt the former group more than it helps the latter and so lower average GPA. The negative average treatment effect of tracking estimated in the section 3 is entirely consistent with this result. Both the theoretical and empirical literature on peer effects (and neighorhood segregation) have emphasized the importance of this complementarity parameter and my results are consistent with the claim that when own and peer characteristics are partially substitutable, tracking reduces mean outcomes.

The inconclusive sign of $\hat{\beta}_{22}$ is also of interest. If this parameter is negative, then GPA is a concave function of peers' high school grades. Average output will therefore be lower with one high and low scoring peer group than with two groups with equal mean high school scores.[23] Unlike $\hat{beta}_{12}$, this parameter does not provide any information about which students benefit most from strong peers. However, it does provide important information about whether mixed or tracked peer groups maximize average output. The role of convexity or concavity in peer effects models has received relatively little attention in the empirical literature but is emphasized in theoretical and methodological work by Benabou (1996) and Graham, Imbens, and Ridder (2011). If both $\hat{\beta}_{12}$ and $\hat{\beta}_{22}$ were negative, the negative average effect of tracking would have been predictable using cross-dormitory variation. Given the ambiguous sign of $\hat{\beta}_{22}$, out-of-sample prediction might have led to incorrect conclusions.

Even if estimates of the two key parameters were consistently negative, predicting the negative

[22] I use a leave-out-one-cluster cross validation scheme that allows for possible dependence of the error structures within dormitry-year clusters. The quadratic model is also the specification that minimizes the Bayesian Information Criterion for this model.

[23] This follows from Jensen's inequality: for a concave function $g(\cdot)$, $\frac{1}{2}g(\mu_{low}) + \frac{1}{2}g(\mu_{high}) \leq g\left(\frac{\mu_{low}+\mu_{high}}{2}\right)$.

effect of tracking would have required an ambitious out of sample extrapolation. Figure 7 shows the density of dormitory-level mean high school grades under tracking and random assignment. Tracking generates a much more dispersed distribution of means ranging from -0.75 to 1.25 standard deviations, while those under random assignment range from -0.25 to 0.65. A similar problem applies to other dormitory-level statistics such as the variance and proportion of students at different percentiles of the high school grade distribution. This figure highlights the limitation of existing methods for predicting the effect of changes in group assignment policies or inferring optimal assignment policies (Bhattacharya, 2009; Graham, Imbens, and Ridder, 2011). These method cannot predict out of sample and so cannot speak to the effect of changing the policies that create peer groups that are not observed under the status quo assignment policy. This may account for the puzzling result in Carrell, Sacerdote, and West (2012), where a change in group assignment policy reduced average outcomes instead of raising them, as estimates based on randomly assigned groups suggested. The "optimal" groups created by the new policy were not observed under the old assignment policy and so their estimated optimality relied on out of sample projection. A similar argument can be applied to aspects of the results in Duflo, Dupas, and Kremer (2011). They estimate variants of the linear-in-means model in equation (5) using students who are randomly assigned to first grade classrooms in Kenyan schools. They interpret this result as evidence that tracking students into classrooms should hurt low-scoring students and help high-scoring students, unless instructors adapt their behavior to respond to the changed classroom composition. However, it is also possible that their estimates sould not be extrapolated out-of-sample to tracked classrooms.

5 Mechanisms

Having established the effects of tracking in section 3 and the effect of marginal changes in peer group composition in section 4, I now consider the mechanisms that might be driving these results. In particular, I suggest that residential peer effects may operate through two conceptually distinct channels: a direct channel, in which peers exert influence through spatial proximity,[24] and an indirect channel, in which spatial proximity influences the formation of social and/or academic networks, which in turn influence student outcomes. The models discussed in section 4 estimated a reduced form combination of these effects. If indirect effects are important, then changes in dormitory assignment policies might change the relationship between spatial proximity and social networks. Estimates that combine direct and indirect effects would then be sensitive to assignment policies and so could not reliably predict the effect of changes in these policies. This further emphasizes the importance of explicit cross-policy comparisons, even if the out-of-sample extrapolation problem were not present.

I present evidence in this section that direct effects alone cannot explain the estimated peer effects. Specifically, I estimate an augmented version of the reduced-form linear-in-means model that

[24]For example, noisy neighbors in the dormitory might impair students' ability to study, even if no direct interaction occurs between the students.

allows peer effects to differ across race groups:

$$GPA_{i,g,t} = \psi_0 + \psi_1 HS_{i,g,t} + \psi_2 \overline{HS}_{g,t} + \psi_3 \overline{HS}_{g,-t} + \nu_{i,g,t}. \quad (8)$$

where $\overline{HS}_{g,t}$ is the average high school grade for students in dormitory g of the same race as student i and $\overline{HS}_{g,-t}$ is the average for students of other races. Table 11 reports the results of estimating this model for dormitory students in the random assignment period and suggest that peer effects occur almost entirely within race groups. Given the salience of race in contemporary South Africa, social and study networks may be strongly correlated with race groups. If this is true, the near-zero cross-race peer effects are evidence that spatial proximity (measured by assignment to the same dormitory) alone does not predict academic outcomes. Instead, the strong within-race peer effects suggest that spatial proximity matters by influencing students' networks, and hence their academic outcomes.

I estimate a similar variant of the linear-in-means model that allows the effect of change in peers' mean high school grades to differ for students in the same and different faculties, such as engineering, humanities, and science. Table 12 shows that peer effects are not stronger within than across faculties. The point estimates for within-faculty effects are in fact consistently smaller than those for cross-faculty effects, though I cannot reject equality of the coefficients in any specification. This suggests that peer effects do not primarily operate through direct academic collaboration (study groups, cooperation on essays, copying problem sets, etc.). The race- and faculty-specific results point to an important role for time allocation and attitudes toward studying, consistent with the conclusions in Stinebrickner and Stinebrickner (2006). The fact that peer effects are almost entirely occurring within race groups also suggests that spatial proximity alone is not sufficient to generate peer effects. Instead, some form of interaction is required and this creates scope for student behavior to respond to changes in group composition or group assignment rules in ways unanticipated by policy makers.

An unusual feature of the assessment system at the University of Cape Town provides additional insight into the time at which peer effects matter. In particular, I show that the treatment effect operates early on in the semester, rather than being concentrated on students' performance in final examinations. Assessment for most courses at most South African universities, takes place in two stages. In the first stage, students are graded on their performance in tests, problem sets, and class discussion, and sometimes their attendance record. Students who perform particularly poorly in this assessment stage may be refused permission to proceed to the second stage of assessment, typically a final examination. I refer to these as "excluded courses" and the transcript data that I use simply shows an exclusion symbol for these courses, rather than a numerical grade.

The results in section 3 assigned zero grades to excluded courses. I can instead recode these as missing courses, calculate students' GPAs using only the non-excluded courses and estimate the treatment effects on this alternative outcome measure. The third and fourth columns of table 8 show that this reduces the average treatment effects on the treated to -0.07. This is approximately half as

large as the treatment effect on the original GPA measure but is still significant. The fifth and sixth columns estimate the treatment effect on the proportion of courses for which students are excluded. This rises from 5 percentage points in the absence of tracking to 8 percentage points under tracking. A large part of the average treatment effect is therefore operating at the extensive margin, through the mechanism of course exclusions.

I also apply the heterogenous treatment effect models from section 3 to this restricted GPA measure. The second panel of figure 4 shows that the treatment effects on GPA calculated from non-excluded courses only are still negative for approximately three quarters of the distribution of high school grades. However, the negative effects are considerably smaller than when using the original inclusive GPA measure and the positive effects in the upper tail are considerably larger (up to 0.3 standard deviations), although they are very imprecisely estimated. The second panel of figure 6 presents the treatment effects at each percentile of the distribution of the restricted GPA measure estimated by the nonlinear difference-in-differences model. These effects are zero for most of the distribution but positive in the upper tail. Together, these figures suggest that course exclusions are an important mechanism through which tracking harms low-achieving students and that after removing these courses, the effects on high achieving students may be marginally positive.

These additional outcome measures provide some important evidence about the mechanisms through which tracking is lowering grades. The fact that course exclusions are based on performance early in the semester suggests that tracking is reducing students' academic aptitude or application from an early stage in the courses, with a particularly deleterious effect on the bottom tail of the distribution. This is more consistent with a model in which peer effects operate from early in the semester than a model in which peers only effect time allocation during intensive study periods just before final examinations.

6 Conclusion

This paper presents a range of evidence showing that the residential tracking policy at the University of Cape Town had a significant negative effect on student GPAs. This effect was driven by particularly large negative effects on the lower tail of the distribution and near-zero effects on the upper tail of the distribution. This policy would be judged as undesirable by almost all standard social welfare functions over the GPA distribution. However, it must be acknowledged that GPA is not a comprehensive measure of student welfare. Other research has noted that peer characteristics and behavior may affect outcomes such as time allocation (Stinebrickner and Stinebrickner, 2006) and attitudes toward diversity (Boisjoly, Duncan, Kremer, Levy, and Eccles, 2006) and such effects cannot be ruled out with the available data.

These results contribute to the broader literatures on peer effects and tracking in education by providing what appears to be the first clean evidence on the effects of noninstructional tracking.

This complements the small literature that cleanly identifies the effect of instructional tracking. For example, Duflo, Dupas, and Kremer (2011) find that although the net effect of instructional tracking is positive, there is suggestive evidence that the direct peer effect of tracking is negative.

While caution should always be taken in generalizing results, three aspects of the findings may be of wider relevance. First, the extremely large negative peer effect on the lower tail suggests that tracking may be a highly regressive policy. Its implementation may be particularly damaging in environments with a highly heterogeneous student population in which the lower tail began with a very weak level of academic preparation. Second, the near-zero effect on the upper tail suggests that high-performing students may be only weakly affected by tracking or respond in ways that offset its effects. It is unclear in this case whether these students' academic performance was directly unaffected by tracking or whether they compensated for a positive effect of tracking by reallocating their time away from studying.

Finally, the direct tests implemented in section 4 showed that in this study, peer effects estimated under random assignment had limited ability to predict the effects of the a change in assignment policy. This difficultly may also apply in other settings. The fact that I find direct evidence that peer effects do not operate through spatial proximity alone shows that caution should be used in extrapolating reduced form estimates that do not explicitly take the behavioral content of peer effects into account. Developing such models may be a fruitful avenue for future research.

References

Ammermueller, A., and J.-S. Pischke (2009): "Peer Effects in European Primary Schools: Evidence from PIRLS," *Journal of Labor Economics*, 27(3), 315–348.

Angrist, J., and K. Lang (2004): "Does school integration generate peer effects? Evidence from Boston's Metco program," *American Economic Review*, 94(5), 1613–1634.

Arnott, R. (1987): "Peer group effects and educational attainment," *Journal of Public Economics*, 32, 287–305.

Athey, S., and G. Imbens (2006): "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, 74(2), 431–497.

Bartholomew, D. (1982): *Stochastic Models for Social Processes*. Wiley, London, 3 edn.

Benabou, R. (1996): "Equity and efficiency in human capital investment: the local connection," *Review of Economic Studies*, 63(2), 237–264.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust difference-in-differences estimates?," *Quarterly Journal of Economics*, 119(1), 249–275.

Betts, J. (2011): "The Economics of Tracking in Education," in *Handbook of the Economics of Education Volume 3*, ed. by E. Hanushek, S. Machin, and L. Woessmann, pp. 341–381. Elsevier.

Bhattacharya, D. (2009): "Inferring Optimal Peer Assignment from Experimental Data," *Journal of the American Statistical Association*, 104(486), 486–500.

Blume, L., W. Brock, S. Durlauf, and Y. Ioannides (2011): "Identification of Social Interactions," in *Handbook of Social Economics Volume 1B*, ed. by J. Benhabib, A. Bisin, and M. Jackson, pp. 853–964. Elsevier.

Boisjoly, J., G. Duncan, M. Kremer, D. Levy, and J. Eccles (2006): "Empathy or antipathy: The impact of diversity," *American Economic Review*, 95(5), 1890–1905.

Brock, W., and S. Durlauf (2011): "Interactions-based Models," in *Handbook of Econometrics Volume 5*, ed. by J. Heckman, and E. Leamer, pp. 3297–3380. Elsevier.

Cameron, C., D. Miller, and J. Gelbach (2008): "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 90(3), 414–427.

Carrell, S., R. Fullerton, and J. West (2009): "Does Your Cohort Matter? Measuring Peer Effects in College Achievement," *Journal of Labor Economics*, 27(3), 439–464.

Carrell, S., B. Sacerdote, and J. West (2012): "From Natural Variation to Optimal Policy? An Unsuccessful Experiment in Using Peer Effects Estimates to Improve Student Outcomes," .

Cooley, J. (2010): "Can Achievement Peer Effect Estimates Inform Policy? A View from Inside the Black Box," .

di Giorgi, G., M. Pellizzari, and S. Redaelli (2010): "Identification of Social Interactions through Partially Overlapping Peer Groups," *American Economic Journal: Applied Economics*, 2(2), 241–275.

DiNardo, J., N. Fortin, and T. Lemiuex (1996): "Labor market institutions and the distribution of wages, 1973 - 1992: A semiparametric approach," *Econometrica*, 64(5), 1001–1044.

Ding, W., and S. Lehrer (2007): "Do peers affect student achievement in china's secondary schools?," *Review of Economics and Statistics*, 89(2), 300–312.

Duflo, E., P. Dupas, and M. Kremer (2011): "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review*, 101(5), 1739–1774.

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

——— (2011): "Peer Effects in Education: A Survey of the Theory and Evidence," in *Handbook of Social Economics Volume 1B*, ed. by J. Benhabib, A. Bisin, and M. Jackson, pp. 1053–1163. Elsevier.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*. Chapman & Hall.

Firpo, S. (2007): "Efficient semiparametric estimation of quantile treatment effects," *Econometrica*, 75(1), 259–276.

——— (2010): "Identification and estimation of distributional impacts of interventions using changes in inequality measures," IZA Discussion Paper 4841.

Foster, G. (2006): "It's not your peers and it's not your friends: Some progress toward understanding the educational peer effect mechanism," *Journal of Public Economics*, 90, 1455–1475.

Garlick, R. (2012): "Identification and Estimation of Mobility Treatment Effects," .

Glewwe, P., and M. Kremer (2006): "Schools, teachers and education outcomes in developing countries," in *Handbook of the Economics of Education, Volume 2*, ed. by E. Hanushek, and F. Welch, pp. 945–1017. North-Holland.

Graham, B. (2011): "Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers," in *Handbook of Social Economics Volume 1B*, ed. by J. Benhabib, A. Bisin, and M. Jackson, pp. 965–1052. Elsevier.

Graham, B., G. Imbens, and G. Ridder (2011): "Measuring the Average Outcome and Inequality Effects of Segregation in the Presence of Social Spillovers," Mimeo.

Guryan, J., B. Jacob, E. Klopfer, and J. Groff (2008): "Using Technology to Explore Social Networks and Mechanisms Underlying Peer Effects in Classrooms," *Developmental Psychology*, 44(2), 355–364.

Hahn, J. (1998): "On the role of propensity score in effcient semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–332.

Hanushek, E., J. Kain, J. Markman, and S. Rivkin (2008): "Does Peer Ability Affect Student Achievement?," 18(5), 527–544.

Heckman, J., and R. Robb (1985): "Alternative Methods for Estimating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman, and B. Singer. Cambridge University Press.

Heckman, J., J. Smith, and N. Clements (1997): "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts," *Review of Economic Studies*, 64(4), 487–535.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the propensity score," *Econometrica*, 71(4), 1161–1189.

Horowitz, J. (2001): "The Bootstrap," in *The Handbook of Econometrics Volume 5*, ed. by J. Heckman, and E. Leamer, pp. 3159–3228. Elsevier.

Hoxby, C. (2000): "Peer Effects in the Classroom: Learning from Gender and Race Variation," Working paper 7867, National Bureau of Economic Research.

Hoxby, C., and G. Weingarth (2006): "Taking Race out of the Equation: School Reassignment and the Structure of Peer Effects," .

Hsieh, C.-T., and M. Urquiola (2006): "The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program," *Journal of Public Economics*, 90(8-9), 1477–1503.

Imberman, S., B. Sacerdote, and A. Kugler (2012): "Katrina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees," *American Economic Review*, 102(5), 2048–2082.

Kang, C. (207): "Classroom Peer Effects and Academic Achievement: Quasi-Experimental Evidence from South Korea," *Journal of Urban Economics*, 61, 458–495.

Kling, J., D. Liebman, and L. Katz (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75(1), 83–119.

Lavy, V., and A. Schlosser (2011): "Mechanisms and Impacts of Gender Peer Effects at School," *American Economic Journal: Applied Economics*, 3(2), 1–33.

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

Marmaros, D., and B. Sacerdote (2002): "Peer and Social Networks in Job Search," *European Economic Review*, 46(4-5), 870–879.

——— (2006): "How do Friendships Form?," *Quarterly Journal of Economics*, 121, 79–119.

Pop-Eleches, C., and M. Urquiola (2012): "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, forthcoming.

Robinson, P. (1988): "Root-n consistent semiparametric regression," *Econometrica*, 56(4), 931–954.

Roderick, M., J. Nagaoka, and V. Coca (2009): "College Readiness for All: The Challenge for Urban High Schools," *The Future of Children*, 19(1), 185–210.

Sacerdote, B. (2001): "Peer effects with random assignment: Results for Dartmouth roommates," *Quarterly Journal of Economics*, 116(2), 681–704.

——— (2011): "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?," in *Handbook of the Economics of Education Volume 3*, ed. by E. Hanushek, S. Machin, and L. Woessmann, pp. 249–277. Elsevier.

Sommers, P., and J. Conlisk (1979): "Eigenvalue Immobility Measures for Markov Chains," *Journal of Mathematical Sociology*, 6, 253–276.

Stinebrickner, T., and R. Stinebrickner (2006): "What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds," *Journal of Public Economics*, 90(8/9), 1435–1454.

Vigdor, J., and T. Nechyba (2007): *Peer Effects in North Carolina Public Schools*pp. 73–102, Schools and the Equal Opportunity Problem. MIT Press.

Yatchew, P. (1997): "An elementary estimator of the partial linear model," *Economics Letters*, 57, 135–143.

Zimmerman, D. (2003): "Peer effects in academic outcomes: Evidence from a natural experiment," *Review of Economics and Statistics*, 85(1), 9–23.

Table 1: Changes in baseline demographic and academic characteristics by group

	(1)	(2)	(3)	(4)	(5)
	Dormitory students		Other students		
	Tracked	Randomized	Tracked	Randomized	Second difference
Panel A: High school grades					
Mean grade (standardized)	.17	.2	0	0	-.03
					(.04)
% "A" students	.2	.21	.15	.18	.02
					(.01)
% "D" students	.1	.1	.12	.13	.01
					(.01)
Panel B: Demographic characteristics					
% female	.5	.52	.52	.51	-.03
					(.02)
% black	.5	.52	.12	.12	-.02
					(.01)
% white	.35	.33	.52	.49	0
					(.02)
% other races	.13	.13	.34	.37	.03*
					(.01)
% English-speaking	.59	.56	.85	.86	.05***
					(.01)
% international	.23	.18	.11	.06	0
					(.01)
Panel C: Year of high school graduation					
% eligible for tracking		.027		.033	.006
					(.004)
% "A" students eligible for tracking		.001		.003	.002
					(.003)
% "D" students eligible for tracking		.055		.048	-.007
					(.016)
Panel D: Geographic location of high school					
In Cape Town	.116	.106	.77	.749	-.012
					(.013)
% "A" students in Cape Town	.112	.078	.811	.785	.008
					(.028)
% "D" students in Cape Town	.246	.228	.775	.762	.004
					(.049)

Notes: Standard errors in parentheses estimated using heteroscedasticity-robust covariance matrix. ***, ** and * denote significance at 1%, 5% and 10% levels respectively.

Table 2: Linear difference-in-differences estimates

	(1)	(2)	(3)	(4)
Sample	Full sample		Restricted sample	
Dependent variable	GPA			
ATT of tracking	-.129*	-.107*	-.111***	-.123***
	(.073)	(.061)	(.034)	(.031)
Dormitory fixed effects		×		×
Year fixed effects			×	×
Individual controls			×	×
# dorm-year clusters	60	60	60	60
# dormitory students	7480	7480	6600	6600
# other students	7188	7188	6685	6685

Notes: Standard errors in parentheses are clustered at the dormitory-year level, with non-dormitory students treated as singleton clusters.
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Restricted sample excludes observations with missing individual controls.
Columns (3) and (4) control for student gender, race, language, nationality, a cubic in high school grades and all possible three-way interactions.

Table 3: Linear difference-in-differences estimates within demographic subgroups

	(1)	(2)	(3)	(4)	(5)	(6)
Sample	All students	Female	Male	Black	White	Other races
Dependent variable	GPA					
	All students	Female	Male	Black	White	Other races
ATT of tracking	-.123***	-.116***	-.122**	-.232***	-.138***	.141**
	(.031)	(.039)	(.05)	(.086)	(.043)	(.069)
Dormitory fixed effects	×	×	×	×	×	×
Year fixed effects	×	×	×	×	×	×
Individual controls	×	×	×	×	×	×
# dorm-year clusters	60	36	42	60	58	59
# dormitory students	6600	3368	3232	3234	2428	845
# other students	6685	3459	3226	689	3484	2376

Notes: Standard errors in parentheses are clustered at the dormitory-year level, with non-dormitory students treated as singleton clusters.
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing individual controls are excluded from the sample.
All columns control for student gender, race, language, nationality, a cubic in high school grades and all possible three-way interactions.

Table 4: Linear difference-in-differences estimates by quartiles of high school grades

	(1)	(2)	(3)	(4)	(5)
Sample	All students	Quartiles of high school grades			
		Fourth	Third	Second	First
Dependent variable		GPA			
ATT of tracking	-.123***	-.224***	-.261***	-.157***	-.091*
	(.031)	(.075)	(.058)	(.048)	(.05)
Dormitory fixed effects	×	×	×	×	×
Year fixed effects	×	×	×	×	×
Individual controls	×	×	×	×	×
# dorm-year clusters	60	59	59	58	57
# dormitory students	6600	1298	1301	1475	1811
# other students	6685	1715	1663	1663	1460
F-test for equality of quartile effects					
without controls			3.706**		
with controls and fixed effects			1.85		

Notes: Standard errors in parentheses are clustered at the dormitory-year level, with non-dormitory students treated as singleton clusters.
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing individual controls are excluded from the sample.
All columns control for student gender, race, language, nationality, a cubic in high school grades and all possible three-way interactions.

Table 5: Nonlinear difference-in-differences estimates of average and inequality treatment effects

	(1)	(2)	(3)	(4)	(5)	(6)
Sample	All students			Restricted sample		
	Treated	Counterfactual	Difference	Treated	Counterfactual	Difference
Mean	.052	.182	-.131*	.084	.178	-.095*
			(.072)			(.056)
Interdecile range	2.238	1.877	.361***	2.185	1.910	.275***
			(.106)			(.102)
Interquartile range	1.023	.911	.112**	1.016	.916	.100*
			(.056)			(.056)
Gini coefficient	.097	.079	.018***	.094	.080	.014***
			(.004)			(.004)
Coefficient of variation	1.016	1.010	.006***	1.015	1.010	.005***
			(.001)			(.001)
Individual controls				×	×	×
Dormitory fixed effects				×	×	×

Notes: Standard errors in parentheses estimated from 1000 bootstrap replications, stratifying by period and group, clustering at the dormitory-year level, and treating non-dormitory students as singleton clusters.
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing individual controls are excluded from the restricted sample.
Columns (4), (5), and (6) use propensity score reweighting to control for dormitory fixed effects and student gender, race, language, nationality, a cubic in high school grades and all possible three-way interactions.

Table 6: Transition matrices during each period

		Panel A: Tracking				Panel B: Randomization			
		Quartiles of college GPA				Quartiles of college GPA			
		4	3	1	1	4	3	2	1
Quartiles of high school grades	4	.37	.22	.13	.28	.39	.25	.15	.21
	3	.29	.25	.19	.27	.28	.23	.23	.26
	2	.13	.18	.28	.41	.16	.19	.27	.39
	1	.05	.04	.23	.68	.06	.08	.22	.63

Notes: Quartiles are defined on the full sample of all dormitory and non-dormitory students in each period but transition probabilities are reported for dormitory students only.

Table 7: Treatment effects on mobility measures

	Tracking	Randomization	Difference
Average probability of changing quartile	.813	.83	-.017
			(.021)
Average number of quartiles changed	.921	.912	.009
			(.038)
"Correlation" between high school and college quartile	.547	.566	-.019
			(.035)

Notes: Quartiles are defined on the full sample of all dormitory and non-dormitory students in each period but transition probabilities are reported for dormitory students only.
Standard errors in parentheses estimated from 1000 bootstrap replications, stratifying by period and group, clustering at the dormitory-year level, and treating non-dormitory students as singleton clusters.
***, ** and * denote significance at 1%, 5% and 10% levels respectively.

Table 8: Linear difference-in-differences estimates with alternative outcome measures

	(1)	(2)	(3)	(4)	(5)	(6)
Sample	All	Restricted	All	Restricted	All	Restricted
Dependent variable	GPA		GPA for non-excluded courses only		Excluded credits	
	All grades		Non-excluded grades only		Excluded credits	
ATT of tracking	-.129*	-.123***	-.076	-.068**	.028***	.027***
	(.073)	(.031)	(.089)	(.033)	(.006)	(.005)
Dep var mean					.05	.049
Dormitory fixed effects		×		×		×
Year fixed effects		×		×		×
Individual controls		×		×		×
# dorm-year clusters	60	60	60	60	60	60
# dormitory students	7480	6600	7449	6576	7480	6600
# other students	7188	6685	7043	6559	7188	6685

Notes: Standard errors in parentheses are clustered at the dormitory-year level, with non-dormitory students are treated as singleton clusters.
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing individual controls are excluded from the sample.
Columns 2, 4, & 6 control for student gender, race, language, nationality, a cubic in high school grades and all possible three-way interactions.
Columns 1 & 2 calculate students' GPA treating excluded courses as zero grades.
Columns 3 & 4 calculate students' GPA using only courses from which they were not excluded.
Columns 5 & 6 measure the credit-weighted percentage of courses from which students were excluded.

Table 9: Direct tests for peer effects in the random assignment period

	(1)	(2)	(3)	(4)	(5)	(6)
Sample	Randomly assigned dormitory students					
Dependent variable	GPA					
Own high school GPA	.362***	.332***	.331***	.363***	.333***	.332***
	(.017)	(.016)	(.016)	(.017)	(.016)	(.016)
Mean dorm high school GPA	.241**	.222**	.22**	.35***	.336***	.216**
	(.089)	(.092)	(.098)	(.075)	(.067)	(.099)
Demographic controls		×	×		×	×
Residence fixed effects			×			×
Excluding outlying low-SES dormitory				×	×	×
Adjusted R^2	.213	.236	.248	.215	.239	.248
# observations	3068	3068	3068	3048	3048	3048
# clusters	30	30	30	28	28	28

Notes: Standard errors in parentheses are clustered at the dormitory-year level
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing high school GPA are excluded from the sample.
Columns 2, 3, 5, and 6 control for gender, language, nationality and race.

Table 10: Tests for nonlinear peer effects

Own high school grade	.4***	.374***	.373***
	(.023)	(.022)	(.022)
Own high school grade squared	.137***	.143***	.142***
	(.017)	(.018)	(.018)
Mean dorm high school grade	.221***	.174**	.316***
	(.058)	(.08)	(.145)
Mean dorm high school grade squared	.306***	.273***	-.159
	(.089)	(.096)	(.178)
Own × mean dorm high school grade	-.129**	-.132**	-.132**
	(.063)	(.06)	(.06)
p-value of test against linear model	0	0	0
Demographic controls		×	×
Dormitory controls		×	×
Dormitory fixed effects			×
Adjusted R^2	.244	.272	.278
# observations	3068	3068	3068
# clusters	30	30	30

Notes: Standard errors in parentheses are clustered at the dormitory-year level
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing HS grades are excluded from the sample.
Columns 2 & 3 control for gender, language, nationality and race.
Sample is all dormitory students in the random assignment period.

Table 11: Tests for the relative importance of different peer groups: race differences

	(1)	(2)	(3)	(4)
Sample	Randomly assigned dorm. students			
Dependent variable	GPA			
Own high school GPA	.362***	.331***	.327***	.327***
	(.017)	(.016)	(.017)	(.017)
Mean dorm high school GPA	.241**	.22**		
	(.089)	(.098)		
Mean dorm high school GPA for own race group			.204***	.159*
			(.058)	(.09)
Mean dorm high school GPA for other race groups			-.003	-.051
			(.054)	(.066)
p-value for test of equal effects within and across races			0	.04
Demographic controls		×		×
Residence fixed effects		×		×
Adjusted R^2	.213	.248	.22	.248
# observations	3068	3068	3068	3068
# clusters	30	30	30	30

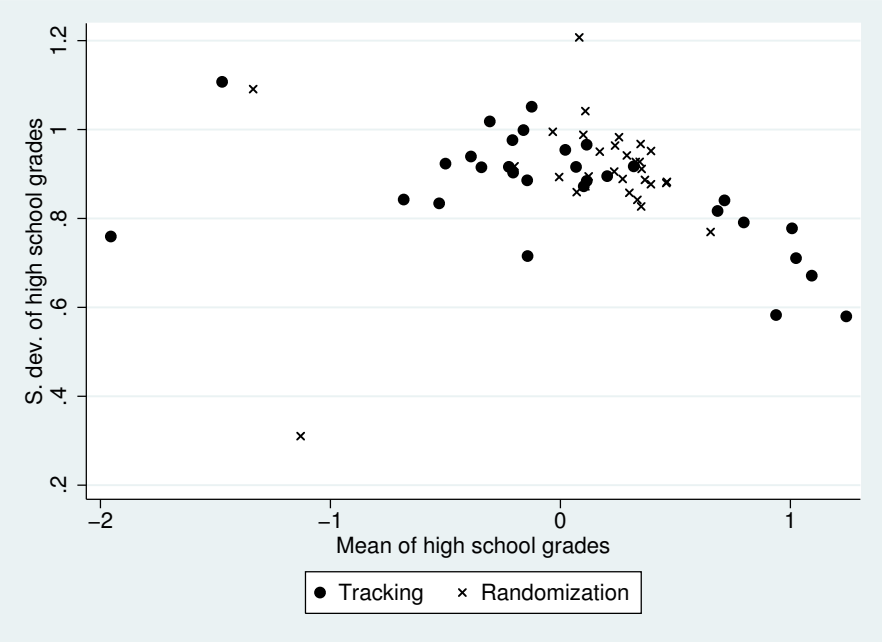
Notes: Standard errors in parentheses are clustered at the dormitory-year level
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing high school GPA are excluded from the sample.

Table 12: Tests for the relative importance of different peer groups: faculty differences

	(1)	(2)	(3)
Sample	Randomly assigned dorm students		
Dependent variable	GPA		
Own HS grades GPA	.345***	.298***	.307***
	(.019)	(.017)	(.016)
Mean dorm HS grades, own faculty	.049	.105**	.066
	(.047)	(.044)	(.062)
Mean dorm HS grades, other faculties	.152*	.171**	.151
	(.072)	(.062)	(.073)
p-value for test of equal effects within and across faculties	.171	.342	.409
Demographic controls		×	×
Dormitory fixed effects		×	×
Faculty fixed effects			×
Adjusted R^2	.202	.239	.251
# observations	3068	3068	3068
# clusters	30	30	30

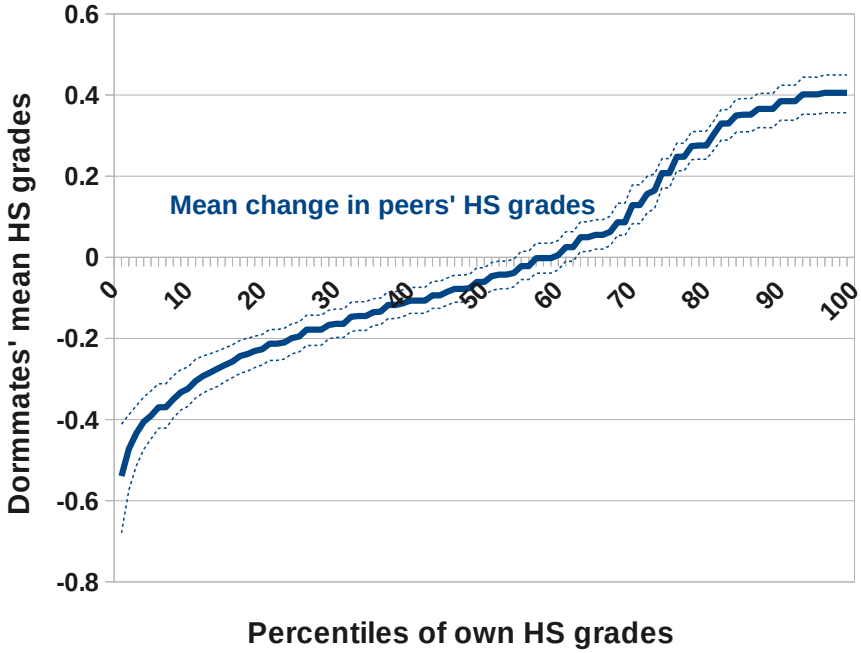
Notes: Standard errors in parentheses are clustered at the dormitory-year level
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing high school GPA are excluded from the sample.

Figure 1: Mean and standard deviation of high school grades by dormitory-year unit



Notes: The four outliers in the left tail are the four yearly observations of a single dormitory with a different assignment rule that typically houses only students from low-income families. It accounts for $\approx 1.4\%$ of the sample and all results are robust to its exclusion.

Figure 2: Change in mean peer high school grades from random assignment to tracking period



Notes: Figure is constructed in four steps. First, I calculate the mean high school grade in each dormitory and assign this to each student as a measure of their peers' baseline ability performance. Second, I estimate a nonparametric regression of peers' baseline performance against own high school grades, separately for the tracking and random assignment period. I use a local linear regression with an Epanechnikov kernel and a plug-in bandwidth following Fan and Gijbels (1996). I allow the bandwidth to differ for the tracking and random assignment periods. Third, I calculate the difference between the two fitted curves at each percentile of the distribution of high school grades. Finally, I construct a 95% confidence interval from the 2.5 and 97.5 percentiles of a nonparametric bootstrap, stratifying by period.

Figure 3: Distribution of high school grades by group in each time period

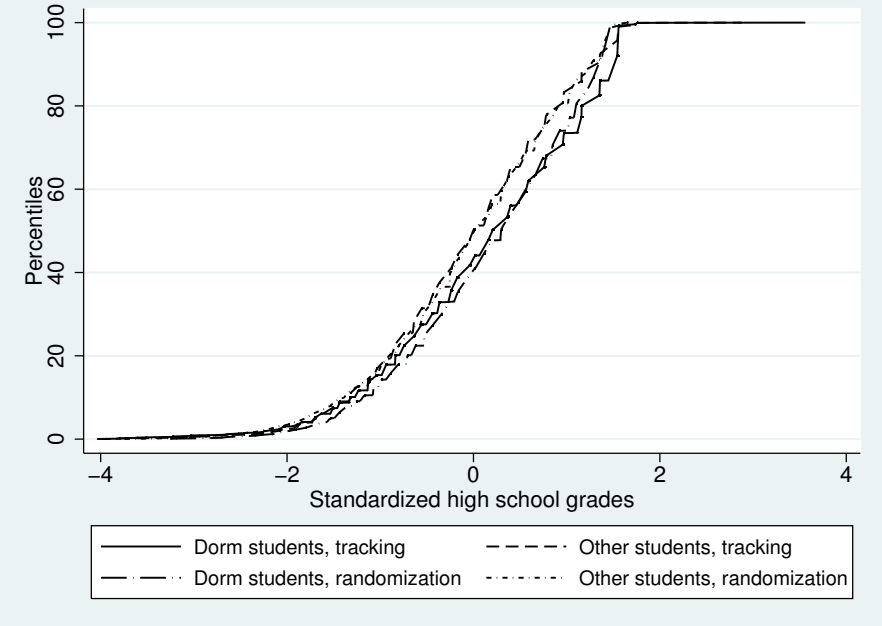
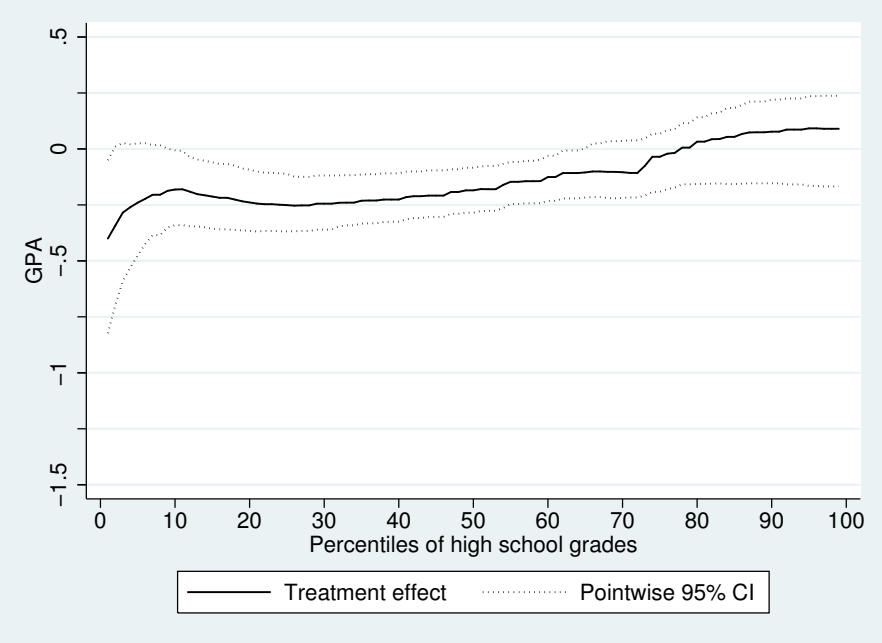
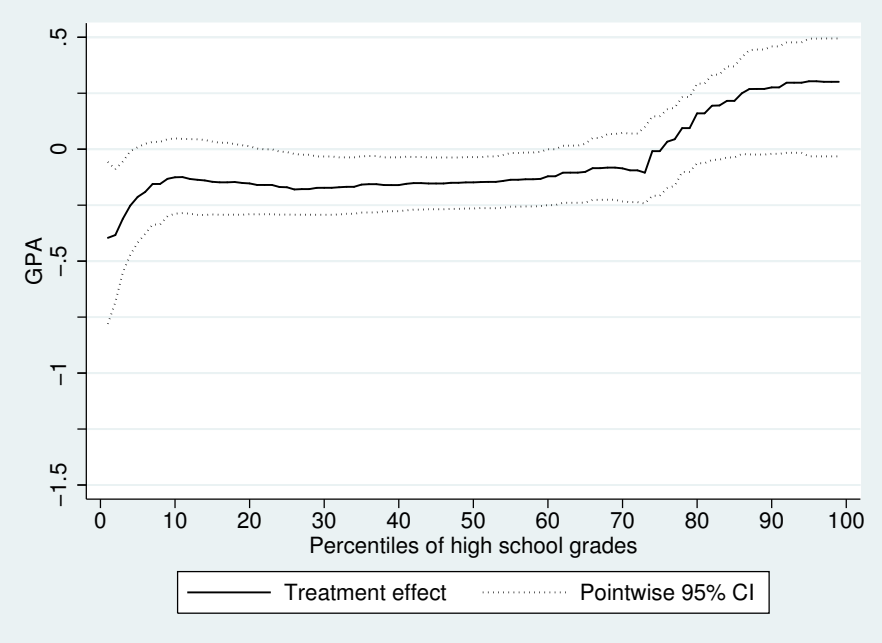


Figure 4: Linear difference-in-differences estimates by percentile of high school grades

Panel A: Using the standard GPA measure



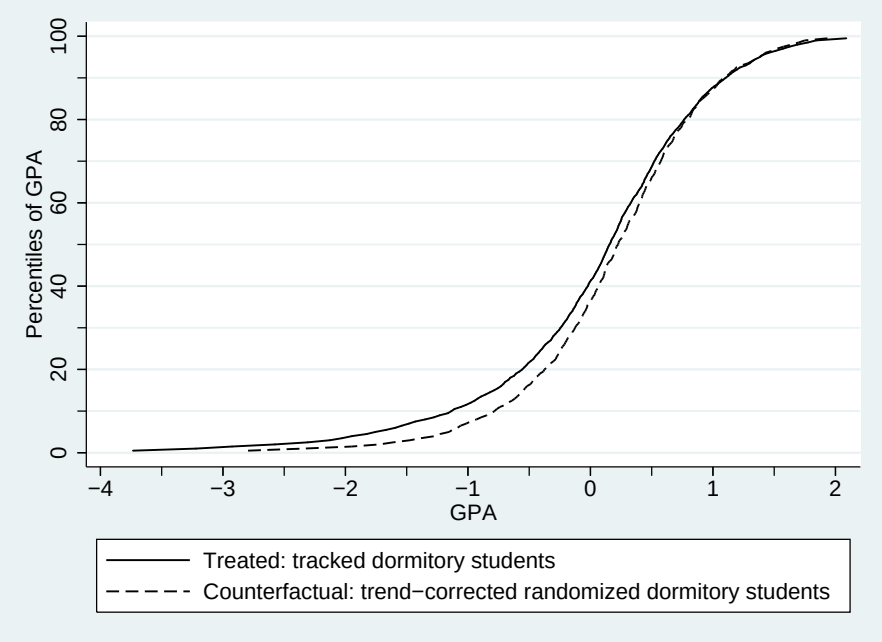
Panel B: Calculating GPA using only non-excluded courses



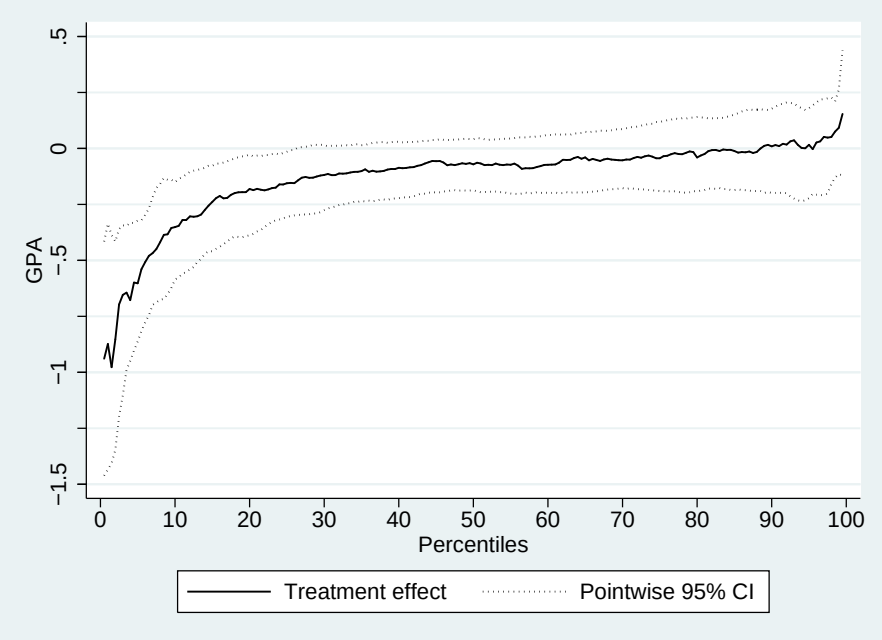
Notes: Using the restricted sample, which excludes observations with missing individual controls.
GPA in the second panel is calculated by treating courses from which students are academically excluded as missing values, rather than zero grades.
Confidence intervals are estimated using a percentile bootstrap with 1000 replications, clustering at the dormitory-year level and stratifying by period and group.

Figure 5: Nonlinear difference-in-differences estimates

Panel A: Distribution of observed outcomes under treatment and counterfactual outcomes

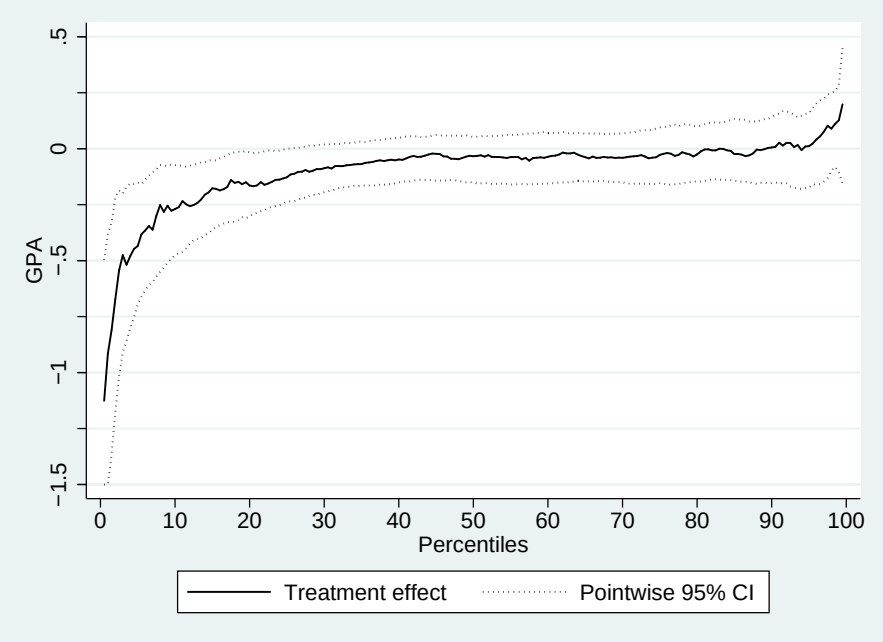


Panel B: Quantile treatment effects

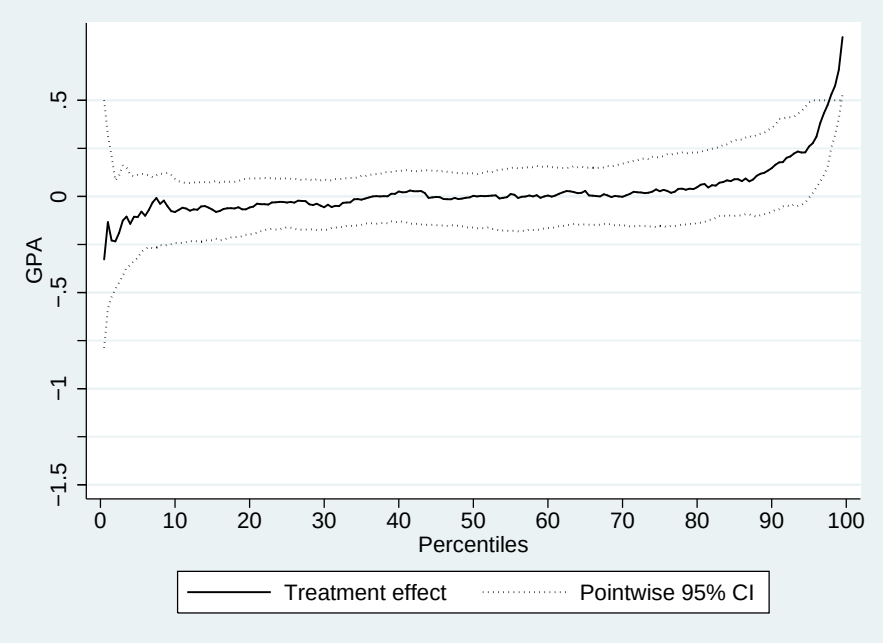


Notes: Using the full sample, which includes observations with missing individual controls. Confidence intervals are estimated using a percentile bootstrap with 1000 replications, clustering at the dormitory-year level and stratifying by period and group.

Figure 6: Reweighted nonlinear difference-in-differences estimates
Panel A: Quantile treatment effects on the standard GPA measure

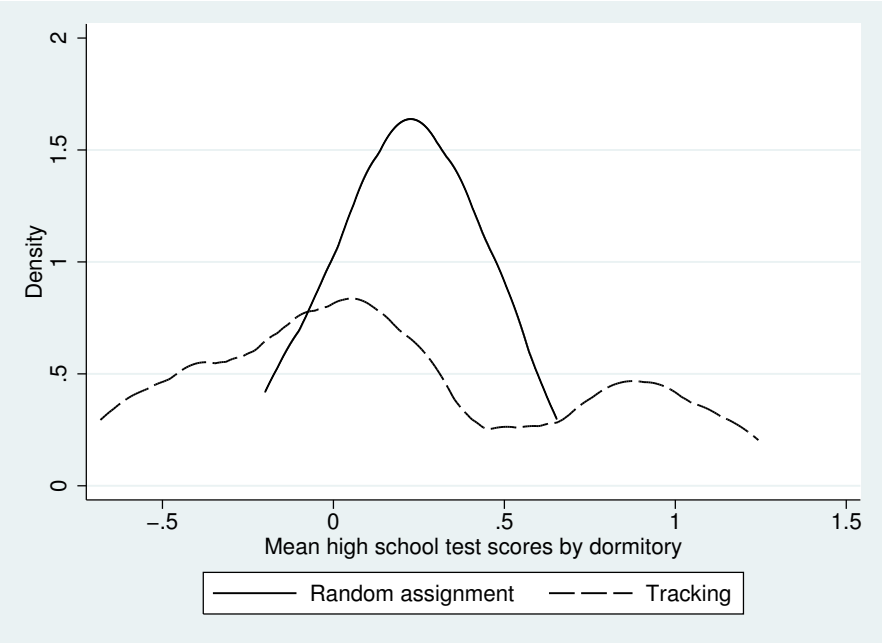


Panel B: Quantile treatment effects on GPA calculated using only non-excluded courses



Notes: Using the restricted sample, which excludes observations with missing individual controls.
Both panels use reweighting to control for student gender, race, language, nationality, a cubic in high school grades, all possible three-way interactions, and dormitory fixed effects.
GPA in the second panel is calculated by treating courses from which students are academically excluded as missing values, rather than zero grades.
Confidence intervals are estimated using a percentile bootstrap with 1000 replications, clustering at the dormitory-year level and stratifying by period and group.

Figure 7: Density of dormitory-level means of high school grades



Notes: Under the random assignment policy, across-dormitory variance is 5% of the total variance. Under the tracking policy, across-dormitory variance is 32% of the total variance.

A Robustness Checks

A.1 Changes in course taking behavior

The body of the paper treated GPAs during the tracking and mixing periods as comparable measures, an assumption that may be problematic if the mix of courses taken is different during the two periods and grading standards vary across courses. I perform three robustness checks to show that the average treatment effects discussed in section 3 are not driven by changes in the mix of courses. In all of these checks, I continue to condition on dormitory, year and student characteristics as in the body of the paper. First, column 2 of table 13 shows that the estimated treatment effect is largely unaffected by the inclusion of faculty fixed effects. Second, columns 3 to 8 report faculty-specific treatment effects. The point estimates are all negative, four of the six are significantly different to zero, and only one is significantly different to the overall treatment effect. The one exception is the law faculty, which is sufficiently small that the extremely large and imprecisely estimated negative point estimate should be treated with caution. Third, table 14 estimates treatment effects within the six largest introductory courses at the university: economics, information systems, management, physics for engineers, sociology, and statistics. All treatment effects are negative, four of the six are significantly different to zero, and none of the six are significantly different to the overall treatment effect. These results show that the treatment effect is not driven by changes in course selections but instead occurs within individual courses.

A.2 Instrumental variables estimates based on admission rules

The identification strategy in the bulk of the paper can be characterized as a selection on observables strategy. I claim that after conditioning on students' demographic characteristics and high school grades, the change in GPAs between the tracking and random assignment periods would be identical for dormitory and non-dormitory students if tracking were not in place. If students choose whether to live in dormitories or in private accommodation or if students choose whether to attend the university during the tracking or mixing period in order to maximize their expected GPA, this assumption may be violated.

I argued in section 2 that such selection on unobserved characteristics is implausible, given the limited information available to prospective students about the dormitory allocation policies and the rules that determine which students are admitted to the dormitory system. In this appendix, I explicitly use these admission rules to construct instrumental variables that exogenously affect the probability that students will be in each of the four groups in my analysis (dormitory and non-dormitory students, during the tracking and random assignment periods).

Specifically, I use the time of high school graduation as an instrument for the period that students enter the university: the instrument takes the value one for students who graduated in 2004 or earlier and the tracking indicator equals one for students who enrolled in the university in 2005 or

earlier.[25] I use the location of each student's high school as an instrument for whether the student enters the dormitory system: the instrument takes the value one for students who attended a high school outside Cape Town and the dormitory indicator equals one for students living in a dormitory. I use the interaction of these instruments as an instrument for the treatment indicator, which equals the interaction of dormitory and the tracking indicators.

These instruments are strongly correlated with the group indicators: the first stage coefficients for the tracking instrument, dormitory instrument, and interaction are 0.927 (standard error 0.006), 0.647 (0.014), and 0.706 (0.031) respectively. The exclusion restriction for the instrumental variables difference-in-differences model is that any direct effect of the year of high school graduation on GPA (*i.e.* any effect not operating through the dormitory assignment policy) is identical for dormitory and non-dormitory students and that any direct effect of whether the student attended a high school inside or outside Cape Town does not change from the tracking to the mixing period. This restriction does not rule out time trends in student characteristics or differences in school quality but requires that these are, respectively, equal for high schools inside and outside Cape Town and unchanging through time.

Table 15 shows that the instrumented treatment effects are similar to those estimated estimated without the instrumental variables correction. They are, however, relatively imprecisely estimated. This may in part reflect a limitation in the "high school in or outside Cape Town" instrument. Students are admitted to the dormitory system based on their home address and the address of the high school that they attend is an imperfect proxy, as some students attend schools far from their home and live in school hostels or dormitories.[26] This means that the instrument suffers from non-classical measurement error. Preliminary derivations suggest that measurement error-induced bias increases the absolute value of the estimated treatment effect.

Given this concern, caution should be exercised in interpreting the instrumental variables estimates. They provide reassuring evidence that the treatment effects reported in the body of the paper are not driven by selection on unobservables but should not be viewed as a definitive refutation of this possibility. The next appendix therefore develops a further sensitivity analysis that demonstrates the robustness of the least squares treatment estimates to at least some forms of selection on unobservables.

A.3 Sensitivity analysis for potential violations of the identifying assumptions

The difference-in-differences model used in the bulk of the paper yields consistent estimators of the average treatment effects of tracking on the treated under certain assumptions on the structure of any unobserved characteristics that influence GPA. In particular, the linear model assumes that the mean change in dormitory students' unobserved characteristics between the two periods was equal to the mean change in non-dormitory students' unobserved characteristics. This assumption is fundamentally

[25] South Africa uses a January to December school year, so this means that students who graduate in December 2004 or earlier are defined as eligible for the tracking period, which included the academic year starting in January 2005.

[26] I do not observe the home address for the majority of students and so cannot use this information directly.

untestable but the discussion and data presented in section 2 suggest that it is reasonable in this context. Furthermore, the fact that the estimated treatment effects are highly robust to controls for observed student characteristics suggest that there are no substantial differences in unobserved characteristics that are correlated with the observed characteristics.

However, there may still be concerns that there exist some unobserved characteristics that satisfy three conditions: they are correlated with GPA, their distribution violates the "equal trends" assumption above, and they are weakly correlated or uncorrelated with the observed characteristics. Here I propose a simple model of this form of selection and show what it implies for the estimated value of the treatment effect. Consider the possibility that GPAs are generated by the model

$$\begin{aligned} GPA &= \delta_{11}DT + \delta_{10}D(1-T) + \delta_{01}(1-D)T + \delta_{00}(1-D)(1-T) + Z\gamma + \epsilon \\ &\equiv \delta G + Z\gamma + \epsilon \end{aligned} \tag{9}$$

where $GPA_{n\times 1}$ is a vector of outcomes, $G_{n\times 4}$ is a matrix indicating group membership (dormitory in tracking period, dormitory in random assignment period, non-dormitory in tracking period, non-dormitory in random assignment period), Z is an unobserved student characteristic, and $\epsilon_{n\times 1}$ is a mean-zero error term uncorrelated with GPA, G, and Z. To reduce the dimension of the problem, I assume that $Z \in \{-1, 1\}$ with $\mathbb{E}[Z] = 0$, $\mathbb{E}[Z|D = 0, T = 1] = \mathbb{E}[Z|D = 0, T = 0] = 0$,[27] and $\mathbb{E}[Z|D = 1, T = 1] = \rho = -\mathbb{E}[Z|D = 1, T = 0]$. Substantively, I interpret Z as "academic orientation," and consider two possible forms of selection:

- Students with high academic orientation ($Z = 1$) prefer to live in dormitories under tracking than under random assignment because this exposes them to similar peers who are also focused on academic performance, so $\rho > 0$.
- Students with low academic orientation ($Z = -1$) prefer to live in dormitories under tracking than under random assignment because this exposes them to similar peers who are also more interested in social and leisure activities, so $\rho < 0$.

This might arise because the set of students who apply to the university differs in their values of Z between the two periods, not necessarily because the same students are being redistributed between the two periods.

As Z is unobserved, the misspecified GPA model is $GPA = \delta G + \epsilon$. Estimating this by least

[27] The assumption that the distribution of Z is identical for non-dormitory students across the two periods simplifies the resultant algebra but can be relaxed without altering the conclusions.

squares yields

$$\begin{pmatrix} \hat{\delta}_{11} \\ \hat{\delta}_{10} \\ \hat{\delta}_{01} \\ \hat{\delta}_{00} \end{pmatrix} \longrightarrow_p \begin{pmatrix} \delta_{11} + \gamma\rho\frac{\mu_{11}}{\sigma_{11}^2} \\ \delta_{10} - \gamma\rho\frac{\mu_{11}}{\sigma_{11}^2} \\ \delta_{01} \\ \delta_{00} \end{pmatrix}$$

where μ_{dt} and σ_{dt}^2 are the mean and variance respectively of the indicator variable $\mathbf{1}\{D = d, T = t\}$. The difference-in-differences test statistic is therefore

$$\hat{\tau} = \hat{\delta}_{11} - \hat{\delta}_{10} - \hat{\delta}_{01} + \hat{\delta}_{00} \longrightarrow_p \tau + 2\gamma\rho\left(\frac{\mu_{11}}{\sigma_{11}^2} + \frac{\mu_{10}}{\sigma_{10}^2}\right). \tag{10}$$

The test statistic is upward biased when academically orientated students are more common in the tracking period ($\rho > 0$) and downward biased when they are less common ($\rho > 0$). Furthermore, the means and variances can be replaced by sample analogues and a "bias-corrected" estimator[28]

$$\hat{\tau}^{BC} = \hat{\tau} - 2\gamma\rho\left(\frac{\hat{\mu}_{11}}{\hat{\sigma}_{11}^2} + \frac{\hat{\mu}_{10}}{\hat{\sigma}_{10}^2}\right). \tag{11}$$

can be constructed for any hypothesized values of ρ and γ.

I use the observed covariates to calibrate plausible values for these parameters. In particular, I note that if the observed binary characteristics in table 1 are transformed so that $X \in \{-1, 1\}$, the largest value of $|\rho|$ is 0.095 (for language). Similarly, the value of γ can be chosen to match the strength of the association between GPA and selected student demographic characteristics. For example, the value of γ implied by the difference between black and white GPAs is 0.23.

Figure 8 plots the value of the bias-corrected treatment effect for selected values of γ and for all $-0.2 \leq \rho \leq 0.2$. The top panel shows that if the unobserved characteristic differs between the tracking and mixing periods as much as the "most different" observed characteristic and is "as important" a determinant of GPA as race, the bias-corrected treatment effect is approximately -0.07 standard deviations (if academically orientated students select out of tracking) or -0.17 standard deviations (if academically orientated students select into tracking).[29]

The second panel repeats the same analysis under the assumption that the true GPA model is $GPA = \delta G + X\beta + Z\gamma + \epsilon$, the misspecified model is $GPA = \delta G + X\beta + \epsilon$, and $\mathbb{E}[Z|X] = 0$. The last assumption is not necessary but it again simplifies the algebra and corresponds to the "worst-case" selection problem, as the bias arising from an unobserved Z that is correlated with X will be reduced by controlling for X. I include in X students' gender, language, nationality, and race, a cubic in

[28]This estimator is actually still biased for τ because $\hat{\mu}_d t/\hat{\sigma}_{dt}^2$ is a consistent but biased estimator of $\mu_d t/\sigma_{dt}^2$. The bias arising from the nonlinearity of the statistic can be reduced using a higher order Taylor series approximation. This correction makes no difference to my results so I omit it for expositional clarity.

[29]Inference on these point estimates should be performed with a degree of caution, as the omission of Z from the GPA model will also inflate the estimated standard errors of the coefficients.

high school grades, all possible threeway interactions and dormitory fixed effects.[30] Conditioning on X marginally attenuates the bias–corrected treatment effects but the difference is negligible, which is consistent with the limited impact on the treatment effect of controling for individual and dormitory characteristics throughout the paper.

This analysis suggests that an implausibly large degree of selection on unobserved characteristics is necessary to explain the estimated treatment effect. This, combined with the results of the previous appendix, strongly suggests that the treatment effect is not driven by violations of the identifying assumptions laid out in section 2.

A.4 Bootstrap-based test procedures

My data is characterized by a natural group structure, as is almost inevitable in any study of peer effects. Dormitory students live together and there may be concerns that the error terms are correlated within these buildings. The main body of the paper addresses this problem using the so called "cluster-robust variance estimator" that generalizes the Eicker-White heteroscedaticity-robust variance estimator. This estimator has the desirable feature of imposing no structure on the nature of the error correlations within dormitory-year clusters[31] but it converges with the square of the number of clusters, rather than the number of students. Bertrand, Duflo, and Mullainathan (2004) note that this may lead to tests with sizes far lower than their nominal rates when the number of clusters is small. Cameron, Miller, and Gelbach (2008) discuss several remedies for this problem and particularly recommend tests based on either cluster or wild cluster bootstraps applied to pivotal statistics.[32]

Table 16 repeats the analyses in table 2, but implements the test that each treatment effect equals zero using four different strategies: the default cluster-robust variance estimator, a cluster bootstrap estimate of the treatment effect's standard error, a cluster bootstrap approximation of the distribution of the test statistic, and a wild cluster bootstrap approximation of the distribution of the test statistic.[33] Panel A shows that the p-values on the test of zero treatment effect are highly robust across different testing procedures. In most cases, the models that do not control for covariates generate estimates that are marginally significant ($0.014 \leq p \leq 0.187$), while those that control for covariates generate highly significant estimates ($0 \leq p \leq 0.007$). Given this robustness, the use of

[30] The previous bias calculation can be readily extended to control for X using standard partitioned regression results. The only substantive change is that γ is now interpreted as the relationship between GPA and Z conditional on X and $\hat{\tau}$ is obtained from a regression of $M_X(GPA)$ on $M_X G$, where $M_X = I - X(X'X)^{-1}X$.

[31] In contrast, both feasible generalized least squares estimators and Moulton-style corrections to ordinary least squares estimators assume that the group and individual components of the error structure are additively separable. My inferences are robust to using both of these estimators but as there is no theory-driven reason to assume additive separability, I do not report these results.

[32] Pivotal statistics are those whose asymptotic distribution does not depend on unknown parameters. This includes most conventional test statistics, as test statistics for single and multiple hypotheses are asymptotically $\mathcal{N}(0,1)$ and $\chi^2(k)$ respectively under standard assumptions. However, this excludes most parameter and standard error estimators, as these are centered around their true but unknown values.

[33] See Cameron, Miller, and Gelbach (2008) for details on the implementation of these various procedures. I use 1000 replications of a stratified cluster bootstrap that resamples dormitory-year clusters (with probability proportional to size) and individual non-dormitory students after stratifying by year. I implement the bootstrap t-statistic procedures with the relevant null hypothesis imposed.

analytical standard errors as a default is relatively innocuous. Furthermore, the third and fourth testing strategies apply only to test statistics and do not yield valid standard error estimates, limiting the ability of the reader to perform tests other than those reported by the author or mentally construct confidence intervals.

Panel B of table 16 reports the results of implenting the same procedures but allowing for error correlation at the *dormitory level*, rather than *dormitory-year level*. Bertrand, Duflo, and Mullainathan (2004) recommend this approach for most difference-in-differences designs that rely on repeated cross-sections as cluster-specific shocks may persist through time. This concern is arguably less relevant in my application because my sample consists of only first year students, so no student appears in the sample in multiple years. This consideration is relevant only to the extent that a shock affecting first year students in dormitory d in year t continues to affect these students in year $t+1$ and hence affects the new cohort of first year students in dormitory d. My inferences are again robust to this more conservative inference procedure: the models that control for covariates still yield highly significant estimates.

B Reweighted Nonlinear Difference-in-differences Model

Given the limited use of the Athey-Imbens nonlinear difference-in-differences model in the applied literature, this appendix provides an overview of the model and the reweighting extension that I propose.

Begin by defining T as an indicator variable equal to one in the tracking period and zero in the mixing period and D as an indicator variable equal to one for dormitory students and zero for other students. Formally, the model is identified under three assumptions:

(A1) GPA in the absence of tracking is strictly continuous and generated by the model $GPA = h(U, T)$, which is monotone in the unobserved scalar U. Note that the function h need not be known and that GPA does not directly depend on D.

(A2) The distribution of the unobserved characteristic remains constant through time for each group, in this case dormitory and non-dormitory students: $U \perp T|D$.

(A3) The support of dormitory students' GPAs is contained in the support of non-dormitory students' GPAs: $supp(GPA|D = 1) \subseteq supp(GPA|D = 0)$.

These assumptions are sufficient to identify the counterfactual distribution of dormitory students' GPAs in the tracking period *in the absence of tracking*, denoted by $F^{CF}_{GPA|D=1,T=1}(\cdot)$. These are the outcomes that the treatment group would have experienced in the treatment period if treatment had not been applied. I follow Athey and Imbens in assuming that the distribution of GPA is strictly continuous and has no mass points, so the cumulative distribution function is invertible.

To understand the identification proof, begin by considering the observed distribution of dormitory students' GPA in the random assignment period, $F_{GPA|D=1,T=0}(\cdot)$. Under assumption $A1$, GPA can be written as $g = h(U,T)$ and so depends only on the unobservable U and the time period. Under assumption $A2$, the distribution of U does not change through time, so the difference between $F^{CF}_{GPA|D=1,T=1}(\cdot)$ and $F_{GPA|D=1,T=0}(\cdot)$ is due entirely to the change from $T=0$ to $T=1$ in h. If h were known, it would be straightforward to evaluate this change. However, the function is unknown and so an indirect argument must be used to construct a mapping from $g = h(u,0)$ to $\tilde{g} = h(u,1)$ for each possible realization u of U.

1. Let A and B denote two students in the random assignment period with GPA g and assume that A is a dormitory student and B is a non-dormitory student. As h does not depend directly whether students living in a dormitory or note, A and B must have the same value of the unobservable u. The common support assumption $A3$ ensures that every dormitory student in the mixing period has an appropriate comparison non-dormitory student with the same GPA.

2. Let $q = F_{GPA|D=0,T=0}(g)$ denote B's location in the distribution of non-dormitory students in the random assignment period and let C be a non-dormitory student in the tracking period with the same location q in her distribution $F_{GPA|D=0,T=1}(\cdot)$. As h is monotone in u and the distribution of U does not change through time, B and C's equal locations imply that they have the same value of the unobservable u. Their GPAs differ only due to the change from $T=0$ to $T=1$ in $h(\cdot,T)$. Let $\tilde{g} = F^{-1}_{GPA|D=0,T=1}(q) = F^{-1}_{GPA|D=0,T=1}\left(F_{GPA|D=0,T=0}(g)\right)$ denote C's GPA.

3. In the absence of tracking, the same analysis could have been applied to dormitory students. Hence, a dormitory student in the tracking period with $U = u$ would have a GPA of $\tilde{g} = F^{-1}_{GPA|D=0,T=1}(q) = F^{-1}_{GPA|D=0,T=1}\left(F_{GPA|D=0,T=0}(g)\right)$. Denote this student by D.

4. As A and D, the two dormitory students in different periods, have the same values of the unobservable, they would have the same locations in their respective distributions, so $F^{CF}_{GPA|D=1,T=1}(\tilde{g}) = F_{GPA|D=1,T=0}(g)$.

Combining the results in the third and fourth bullet points yields the counterfactual distribution of GPAs for dormitory students in the tracking period:

$$F^{CF}_{GPA|D=1,T=1}(\tilde{g}) = F_{GPA|D=1,T=0}\left(F^{-1}_{GPA|D=0,T=0}\left(F_{GPA|D=0,T=1}(\tilde{g})\right)\right) \tag{12}$$

The quantile treated effect on the treated at quantile q is defined as the horizontal distance between the observed and counterfactual distributions:

$$\Delta(q) = F^{-1}_{GPA|D=1,T=1}(q) - F^{CF,-1}_{GPA|D=1,T=1}(q) \tag{13}$$

Intuitively, the first assumption $A1$ imposes sufficient structure on the data generating process to allow us to compare students' GPAs across groups and their locations in the GPA distribution through time. The second assumption ensures that changes through time in the GPA distribution are due to time changes, which are common to both groups, not changes in the distribution of unobserved characteristics.

Note that the model imposes structure only on the data generating process for GPA in the absence of treatment (tracking) and remains agnostic regarding the data generating process for GPA under tracking. This identifies the counterfactual distribution of GPAs for dormitory students under tracking if tracking had not been implemented but provides no information about the counterfactual distribution of GPAs for non-dormitory students if they had been exposed to tracking. The model therefore identifies the *treatment effect on the treated*, in the framework of Heckman and Robb (1985).

Athey and Imbens also consider a "quantile difference-in-differences" model that computes second differences between groups in quantiles of the outcome: $F^{-1}_{GPA|D=1,T=1}(q) - F^{-1}_{GPA|D=1,T=0}(q) - F^{-1}_{GPA|D=0,T=1}(q) - F^{-1}_{GPA|D=0,T=0}(q)$. This is perhaps a more intuitive approach but it requires the stronger assumption that the distribution of the unobservables is identical for both groups in both time periods. This stronger assumption permits a direct comparison of GPAs through time and across groups, rather than using the location comparison through time. This stronger assumption is clearly inappropriate in my application (see table 1) but my results are robust to using this alternative model.

The original paper only briefly considers the role of observed student characteristics and suggests two somewhat *ad hoc* means of controling for these characteristics. The first suggestion is to implement the model separately for specific values of the covariates (for example, for male and female students). However, this is clearly infeasible with many discrete covariates or with continuous covariates. The second suggestion is to "residualize" GPA by regressing it on the covariates and the applying the model to the residuals from this regression.

I instead use a reweighting scheme, implemented separately for dormitory and non-dormitory students. Specifically, I define the reweighted counterfactual distribution as

$$F^{RW,CF}_{GPA^{11}}(g) = F_{GPA^{10}_{\omega}}\left(F^{-1}_{GPA^{00}_{\omega}}\left(F_{GPA^{01}}\left(g\right)\right)\right) \tag{14}$$

where $F_{GPA^{d0}_{\omega}}(\cdot)$ is the distribution function of $GPA \times Pr(T=1|D=d,X)/Pr(T=0|D=d,X)$. Intuitively, this scheme assigns high weight to students in the random assignment period whose observed characteristics are similar to those in the tracking period. This is a straightforward extension of the reweighting techniques used in the wage decomposition literature (DiNardo, Fortin, and Lemiuex, 1996) and the program evaluation literature (Hirano, Imbens, and Ridder, 2003). Firpo (2007) lays out the technical assumptions under which the reweighted distribution is consistently estimated by the predicted probabilities from a series logistic regression of T on X. Under these assumptions

$$\hat{\tau}^{QTT}(q) = \hat{F}^{-1}_{GPA^{11}}(q) - \hat{F}^{-1,RW,CF}_{GPA^{11}}(q) \tag{15}$$

is a consistent estimator of the quantile treatment effect on the treated in the *reweighted nonlinear difference-in-differences* model.

An important assumption invoked for consistency of this reweighted estimator is that the propensity score $Pr(T = 1|D = d, X)$ is consistently estimated. Firpo (2007) proposes doing so using a semiparametric logistic model in which T is regressed on a polynomial function of X whose order satisfies certain regularity conditions. Selecting the order of the polynomial is a difficult process and the literature provides relatively little guidance. In practice, I use polynomial orders from 1 to 4, and the choice of this tuning parameter makes little difference to my results.

I implement the estimator in three steps:

1. I regress an indicator for the tracking period on a flexible logistic function of X, separately for each group, and use the predicted probabilities from that regression to construct $\hat{Pr}(T = 1|D = d, X)$ for each student.

2. For each half-percentile of the distribution of GPAs (i.e. quantiles 0.5 to 99.5), I implement equation (14) to construct the reweighted counterfactual distribution of GPAs in the absence of tracking.

3. I then replicate this process 1000 times on bootstrap resamples of the data, clustering at the dormitory-year level and stratifying by group and period, to construct percentile bootstrap confidence intervals for the estimated treatment effect at each of the 199 quantiles from step 2.

Note that I do not attempt to estimate the counterfactual minimum and maximum, as inference on these parameters is known to be highly problematic (Horowitz, 2001).

Table 13: Linear difference-in-differences estimates within faculty/school

Sample	(1) All students	(2) All students	(3) Commerce	(4) Engineering	(5) Law	(6) Medicine	(7) Science	(8) Humanities & Soc. Sci.
Dependent variable	GPA							
ATT of tracking	-.123***	-.137***	-.202***	-.196***	-1.297***	-.022	-.028	-.128*
	(.031)	(.03)	(.057)	(.07)	(.304)	(.093)	(.075)	(.054)
Dormitory fixed effects	×	×	×	×	×	×	×	×
Year fixed effects	×	×	×	×	×	×	×	×
Individual controls	×	×	×	×	×	×	×	×
Faculty fixed effects		×						
# dorm-year clusters	60	60	60	60	42	43	60	60
# dormitory students	6600	6600	2060	1382	124	678	1052	1304
# other students	6685	6685	1952	1004	113	453	824	2339

Notes: Standard errors in parentheses are clustered at the dormitory-year level, with non-dormitory students are treated as singleton clusters.
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing individual controls are excluded from the sample.
All columns control for student gender, race, language, nationality, a cubic in high school grades and all possible three-way interactions.

Table 14: Linear difference-in-differences estimates within entry-level courses

Sample	(1) Economics	(2) Information systems	(3) Management	(4) Physics	(5) Sociology	(6) Statistics
Dependent variable	GPA (with zeros for excluded credits)					
ATT of tracking	-.149***	-.146**	-.184***	-.263***	-.099	-.065
	(.045)	(.071)	(.062)	(.09)	(.092)	(.051)
Dormitory fixed effects	×	×	×	×	×	×
Year fixed effects	×	×	×	×	×	×
Individual controls	×	×	×	×	×	×
# dorm-year clusters	59	56	60	59	55	60
# dormitory students	2160	1174	1322	822	547	1801
# other students	2094	1228	1296	574	969	1978

Notes: Standard errors in parentheses are clustered at the dormitory-year level, with non-dormitory students are treated as singleton clusters.
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Observations with missing individual controls are excluded from the sample.
All columns control for student gender, race, language, nationality, a cubic in high school grades and all possible three-way interactions.

Table 15: Instrumental variables estimates of the average treatment effect of tracking

	(1)	(2)	(3)	(4)
Sample	All students	Non-missing geocodes		
Dependent variable	GPA			
Estimator	OLS	OLS	ITT	2SLS
Avg treatment effect	-.129	-.114	-.104*	.144*
	(.073)	(.074)	(.059)	(.077)
# dorm-year clusters	60	60	60	60
# dormitory students	7480	6187	6187	6187
# other students	7188	6110	6110	6110

Notes: Standard errors in parentheses are clustered at the dormitory-year level, with non-dormitory students treated as singleton clusters.
***, ** and * denote significance at 1%, 5% and 10% levels respectively.
Columns 2 – 4 exclude observations with missing high school data

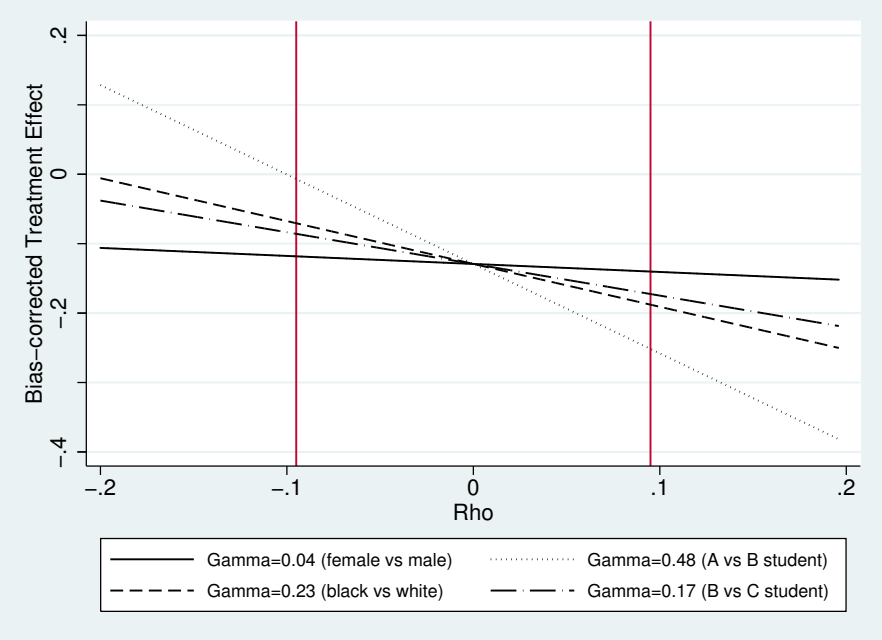
Table 16: Linear difference-in-differences estimates with different test procedures

	(1)	(2)	(3)	(4)
Sample	Full sample		Restricted sample	
Dependent variable	GPA			
ATT of tracking	-.129	-.107	-.113	-.123
Dormitory fixed effects		×		×
Year fixed effects			×	×
Individual controls			×	×
Panel A: p-values for test of zero treatment effect, dormitory-year clusters				
cluster-robust variance estimator	.077	.078	.001	0
cluster bootstrap variance estimator	.079	.187	.001	.001
cluster bootstrap t-statistic	.014	.062	0	0
wild cluster bootstrap t-statistic	.143	.09	.005	.007
Panel B: p-values for test of zero treatment effect, dormitory clusters				
cluster-robust variance estimator	.146	.254	.001	.001
cluster bootstrap variance estimator	.155	.27	.001	.001
cluster bootstrap t-statistic	.098	.191	0	.001
wild cluster bootstrap t-statistic	.266	.363	.014	.012

Notes: Restricted sample excludes observations with missing individual controls.
Columns 3 & 4 control for student gender, race, language, nationality, a cubic in high school grades and all possible three-way interactions.

Figure 8: Linear difference-in-differences estimates' sensitivity to possible selection on unobservables

Panel A: No individual controls



Panel B: Controling for individual and dormitory characteristics